

nature

THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE

FREEDOM OF MOVEMENT

Electrons achieve
perfect tunnelling
from metal to
superconductor

PAGE 344

MICROBIOLOGY

HOW TO SPEAK VIRUS

The surprisingly sociable
world of the bacteriophage

PAGE 290

MATERIALS SCIENCE

INK-FREE PRINTING

Stress-induced colour in
transparent polymers

PAGES 312 & 363

EVOLUTION

SELFISH SEX CHROMOSOMES

Why female mate choice
sometimes puts males at risk

PAGES 311 & 370

NATURE.COM

20 June 2015

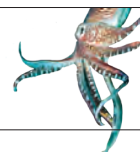
Vol. 519, No. 7563

THIS WEEK

EDITORIALS

DIVERSITY Road to equality programmes paved with good intentions **p.277**

DEVELOPMENT Cuttlefish and spiders use same genes to grow limbs **p.277**



EBOLA Virus spreads from the DRC to Uganda **p.281**

A fresh chance to feed the world

The Food and Agriculture Organization of the United Nations is about to elect a new director-general. The choice will affect the entire globe for years to come.

Charting the future of the world's food production is one of the biggest challenges we face. Feeding the 10 billion mouths expected on the planet by 2050 will be demanding, but should be doable. Much more difficult will be doing it without irreparably damaging the planet.

That's why the election next week of the new director-general of the Food and Agriculture Organization of the United Nations (FAO) is important. The Rome-based intergovernmental body — which has 194 member states — is the pre-eminent international voice on almost every issue touching on food and agriculture. The agency and its leader — who typically serves for two four-year terms — wield considerable influence on global and local policies and play a key part in agricultural research. Member nations must choose the right person for the job.

The FAO emerged from the major disruption to agriculture and food supply caused by the Second World War. Launched in October 1945, its goals were to promote scientific research, share knowledge and collect detailed and comprehensive statistics, with the goal of alleviating hunger, improving nutrition and productivity, and boosting the livelihoods of poor farmers.

The agency's work today spans fisheries, farming and forestry, and is closely entwined with research. The FAO maintains the FAOSTAT database, which contains extensive data on indicators such as crop production, fertilizer use and food security, and which is used by researchers worldwide. In cooperation with the World Health Organization (WHO) and other bodies, the FAO has helped to restructure farming industries and the livestock trade to reduce the risk of emerging infectious diseases. It has scientists, statisticians and other specialists on its staff, collaborates with academic institutions and researchers to inform its policies, and makes science-based recommendations, such as on antibiotic use in farmed animals and the use of genetically modified crops and livestock.

The future challenges for the agency are huge. It must help to steer food production in many different ecosystems to adapt to climate change as well as the extreme weather events that will accompany it. Population increase, rapid urbanization and shifting diets also demand major changes in farming and food production. So the new head will need to show strong and visionary leadership, set a more sustainable agenda to which countries will sign up, and stand up to vested interests from the agricultural and food industries.

The world's food-production systems generate as much as half of the globe's greenhouse gases, erode soils, pollute the environment, and harm biodiversity by destroying natural habitat. Intensive livestock production is one of the worst offenders, yet the billions of people escaping poverty or becoming the new urban middle class tend to eat more meat than before, so livestock production is on track to double by 2050. Producing more food sustainably demands major shifts in policy and practice, as well as innovation, to lessen agriculture's environmental footprint. This will include reducing the use of pesticides, herbicides and artificial fertilizers, and drastically cutting food waste. Around

one-third of food produced for humans currently goes to waste.

The global food system also produces too much highly processed food — contributing to an increase in obesity and diet-related diseases worldwide. There are now more people in the world who are obese than are underweight. The current director-general, José Graziano da Silva, has helped to drive a shift in thinking at the FAO and its member states, away from an excessive focus on producing more food and calories to also promoting access to healthier and more nutritious foods. His successor must continue these efforts.

There are three candidates in the running, down from an original five. They are Qu Dongyu, a former academic policy researcher and China's vice-minister for agriculture and rural affairs; Davit Kirvalidze, an agricultural scientist and former agriculture minister from Georgia; and France's Catherine Geslain-Lanéelle, an engineer in agronomy and former executive director of the European Food Safety Authority. She is the

"Producing food sustainably demands major shifts in policy and practice, as well as innovation."

first candidate to be put forward collectively by the European Union and the first female candidate for the post in the FAO's history.

All the candidates are highly qualified for the job — but it would be naive to think that the winner will be selected on merit alone. The FAO election process — a secret ballot in which every country has a single vote — means that the process is rife with geopolitics

and political horse-trading. China in particular is reported to be unabashedly leveraging its influence and its investments in its massive Belt and Road Initiative, to get votes for its candidate. Securing the top FAO job would be a diplomatic win for China and give it weight in setting global food policy.

The behind-the-scenes bargaining means that there is little incentive for the candidates to publicly discuss and debate their visions for the agency in any depth. That's a shame. The election of the director-general of the WHO used to have the same flaws but has opened up, under pressure. The FAO election must be more transparent in future, with candidates obliged to detail and debate their manifestos. The incoming director-general must make greater transparency and openness across the agency a priority.

Greater transparency might also encourage greater donor confidence. With around 11,500 staff, the FAO is the largest technical UN agency, but its budget — US\$2.6 billion for 2018–19 — has been flat in recent years. The incoming leader must do more to raise the profile of the FAO and make it an attractive investment for member countries and other donors. Otherwise, the agency will end up with its hands largely tied.

The FAO is in the unusual position of being able to help feed the world and, by doing it sustainably, to save the world from catastrophic environmental degradation. In an era in which some countries are turning towards nationalism and rejecting international cooperation, it's even more important that the voting nations select a leader who has these global interests at heart. ■



Unintended consequences of gender-equality plans

Don't let academia's initiatives to advance women become just another way to game the research system, urges Charikleia Tzanakou.

As someone who studies academic organizations and their efforts to encourage equality, I believe that dedicated programmes to address inequalities are essential. I also worry they might not be designed to provide support to the very people who need it. Issues such as race, class and overlapping patterns of discrimination must be considered. So, too, must the way that the measurements used for assessment tend to distort what is being measured.

The rationale for programmes to promote women's representation in science is clear. In the European Union, only 15% of senior academic positions in science are held by women. Numbers are improving (slowly), and gender-equality initiatives deserve some of the credit.

One of the most prominent is the Athena SWAN charter, which was established in the United Kingdom in 2005 and has inspired programmes in other countries. It grants awards to research institutions that perform self-assessments on gender representation and career advancement, and establish well-documented multi-year plans to improve. As of April 2019, more than 160 institutions had collected more than 800 bronze, silver or gold Athena SWAN awards. In 2011, the UK National Institute for Health Research announced that medical schools needed a silver or gold award to be eligible for its funding. Since then, the proportion of applications from relevant departments has increased fourfold.

Initiatives such as Athena SWAN insist on hard data, which are crucial for credibility and accountability. But, too often, these data focus on women as a homogenous group and so overlook intersecting patterns of disadvantage faced by women of colour, early-career researchers and sexual minorities. For example, compared with white women, female academics of colour report limited access to mentoring and higher rates of feeling isolated, excluded, discounted and not belonging.

Another unintended consequence is that women can be penalized by the programmes designed to help them. A 2014 analysis found that women make up more than 70% of Athena SWAN champions, a labour-intensive role that takes time away from their research. My colleague and I found similar patterns in an analysis of 11 institutions with silver SWAN awards: ten teams had more female members than male; eight of the submission teams were led by women. One interviewee described coming in to work on her department's Athena SWAN application on a Sunday and having e-mail conversations with two women at other institutions who were also working on theirs. Some women had been advised that their promotion would depend on attaining an award, even though the department had not provided essential resources or support. That defies the spirit of the awards. These programmes should be about the ability of the department to support equality.

Perhaps the most unfortunate unintended consequence is that achieving gender equality becomes a box-ticking exercise, divorced from the broader goals for a fairer society. A department looks at

gender-equality data not as an opportunity to gain insight and improve the working environment for all, but to present itself in a certain light in order to secure the award; it must assert that inequality is not really that bad within their unit, but that it can make clear improvements. There is a temptation to think more about what can be demonstrated than about what needs to be done.

Gender-equality programmes should be about collaboration and working together to weaken entrenched inequalities in the academic system. How can we make sure that happens? I have some ideas.

First, explicitly consider intersecting patterns of disadvantage. Calculating the percentage of women in various positions is insufficient; it is important to capture experiences of sexual, ethnic and other minorities as well. Fortunately, equality programmes, including Athena SWAN, are broadening reporting requirements around intersectionality and people from gender or sexual minorities, but the focus is still often on simpler, overarching statistics.

Second, the awards should consider more qualitative data about the workplace culture, and make sure that applicants have the resources to support this extra work. Assessments should, ideally, be accompanied by site visits. The award system for Project Juno, a gender-equality and inclusion programme set up by the Institute of Physics in the United Kingdom, offers a visit by evaluators to the research institution and meetings with staff.

Evaluators should look beyond data to find the stories behind career decisions. Women I interviewed told me that they left positions because

they felt that their career progression was blocked. None had reported this to her former employer during the exit interview.

Third, create ways for people to report mistakes and disappointments without jeopardizing their award status. Everyone I interviewed had seen unintended consequences of gender actions (everything from unfair burdens, to less willingness to talk about the most difficult problems, to stronger feelings of hypocrisy and cynicism). A dedicated space in the application form to describe these would reduce the incentive to put a misleading positive spin on reports. There are precedents for helping people to air dirty laundry — reports could be anonymized or shared under strict Chatham-House rules (that is, submitted to a trusted group under strict confidentiality). Most people who start equality and diversity initiatives truly want to learn from mistakes and to help others achieve the same goal.

Most of all, continue to check that the consequences of programmes match their intent. Gender-equality initiatives are laudable, but their drawbacks and insufficiencies should not be ignored. The only way not to suffer from unintended consequences is to be mindful of them. ■

Charikleia Tzanakou co-leads a Horizon 2020 PLOTINA project on academic culture change at the University of Warwick, Coventry, UK. e-mail: charikleia.tzanakou@warwick.ac.uk

**ACHIEVING
GENDER EQUALITY
BECOMES A
BOX-TICKING
EXERCISE.**

SEVEN DAYS

The news in brief

POLICY

Harassment study

Several major US science agencies say they have received few reports of sexual harassment by the researchers whose work they fund, despite studies that have found such behaviour to be pervasive in US academic science. In some cases, agencies have turned to news reports to identify researchers under investigation by universities for such behaviour, according to a report released on 12 June by the US Government Accountability Office. NASA received three sexual-harassment complaints about its grant recipients between 2015 and 2019, the Department of Energy received two, and the Department of Health and Human Services — which includes the National Institutes of Health — received one. The National Science Foundation, which last year began requiring institutions that it funds to report any finding they make regarding sexual harassment by a grant recipient, received 14 complaints.

Genetics law

China has announced a new law restricting the collection and use of genetic resources from people in the country — including biological samples that yield DNA, such as blood, and data gleaned from sequencing them. The law, which goes into effect on 1 July, formalizes restrictions on such activities that have been in place since 1998. Scientists working for foreign organizations will still need to collaborate with a domestic research organization to work with genetic material from Chinese citizens and to take DNA resources outside China, and such



GEORGE LO/SOPA IMAGES/LIGHTROCKET/GETTY

Academics back Hong Kong protesters

Hong Kong's government has suspended an unpopular bill that would make it legal to extradite people to mainland China to stand trial or serve criminal sentences. The 15 June move was a response to huge protests that began more than a week ago. But the bill could still be revived, and protests continue, with critics demanding its withdrawal. More than 1,000 international academics have signed a petition against the

proposed law, saying it would restrict free speech and academic debate, and erode personal freedoms. When the bill was introduced, Hong Kong chief executive Carrie Lam said it was meant to close a loophole that allows criminals to stay in the city, but critics say that mainland China — which does not have Hong Kong's legal protections — would crack down on dissidents living in or visiting the city.

collaborations require the science ministry's approval. Fines for unauthorized collection of genetic resources or data can be up to 10 million yuan (US\$1.4 million).

Advice not wanted

US President Donald Trump has signed an executive order that directs federal agencies, including science agencies, to reduce the number of advisory committees they support by one-third before the end of September. Agencies can no longer create advisory committees unless

the total number of panels across the government is less than 350; currently, about 1,000 advisory committees provide the government with independent input on a plethora of scientific, environmental and health issues, among others. The Trump administration says the policy, enacted on 14 June, will increase government efficiency, but critics say that it will undermine the role of science in government decisions — and note that the administration has moved to restrict the influence

of advisory committees at agencies including the Environmental Protection Agency. "It's no longer death by a thousand cuts," said Gretchen Goldman, an analyst at the Union of Concerned Scientists, an advocacy group in Cambridge, Massachusetts, in a statement. "It's taking a knife to the jugular."

CLIMATE CHANGE

UK climate goal

Scientists have welcomed a commitment by the UK government to reduce net

carbon emissions to zero by 2050. Draft legislation laid out in Parliament on 12 June would tweak the 2008 Climate Change Act, which sets an emissions-reduction target of 80% by 2050, compared with 1990 levels. Climate scientists praised the promise to raise that target to 100%, but they stress that the commitment is just the first step in ending the country's contribution to climate change. "We need to see not only the legislation and the promises, but also action to make those promises real," said Tim Kruger of the University of Oxford, UK. The move follows a May report from the Committee on Climate Change, an independent advisory body, which recommended the net zero target to meet the United Kingdom's commitments to the 2015 Paris agreement. The draft law must now be approved by Parliament.

SPACE

China space station

China has selected nine scientific experiments to fly on its next major space station, scheduled for completion in 2022. The China Manned Space Agency chose the projects from 42 hopefuls, in a process organized with the United



Nations. China's existing space laboratory, Tiangong-2, also hosts experiments, but the new outpost — currently known as the China Space Station — will be bigger and longer-lasting (pictured, Chinese astronauts in training). The projects cover topics from astronomy and biology to the behaviour of fluids and fire in space. A European experiment will study how microgravity and radiation in space affect the mutation of human DNA, and a Russian–Indian experiment will study star formation by mapping the cosmos in ultraviolet light. Scientists working on the projects hail from nations that have space programmes, as well as countries such as Kenya, Mexico and Peru — the result of an effort to encourage participation from low- and middle-income countries. But the United States is not involved in any of the experiments. Since

2011, NASA researchers have been banned from collaborating with China without congressional approval.

HEALTH

Ebola spread

A five-year-old boy in Uganda who was infected with Ebola died on 12 June, according to the Ugandan ministry of health. The child was the first person to be diagnosed with the illness in the country since an outbreak of the virus emerged in the neighbouring Democratic Republic of the Congo (DRC) ten months ago. He had travelled from the DRC to Uganda with his family on 10 June, the World Health Organization (WHO) said. The Ugandan health ministry confirmed that the boy had tested positive for Ebola, as had his three-year-old brother and his grandmother, who also died.

The WHO and Uganda's health ministry have dispatched rapid-response teams to try to limit Ebola's spread in the country. The Ebola outbreak in the DRC is now the second deadliest on record, with roughly 1,400 deaths. There had been more than 2,100 confirmed and probable cases of Ebola there as of 16 June, according to the WHO. See page 283 for more.

EVENTS

Satellites lift off

Three Canadian radar satellites that will monitor environmental changes in Canada and the Arctic launched on 12 June from Vandenberg Air Force Base in California. The long-awaited RADARSAT Constellation Mission will image areas of interest up to four times a day and will be able to track ships across the oceans in the north. The Canadian government, which will run the trio of spacecraft, approved the project in 2008, but the mission ran over budget and behind schedule. The first two RADARSAT satellites, which launched in 1995 and 2007, enabled the first high-resolution mapping the whole of Antarctica, among other things.

TREND WATCH

People living in wealthy countries are more likely to question the safety of vaccines than are those living in poorer ones, according to a worldwide survey on public attitudes to science and health.

More than 140,000 people from around 140 countries gave their views on the safety and efficacy of vaccinations as part of an almost two-year project funded by London-based biomedical research charity Wellcome.

The results of the Wellcome Global Monitor, published on 19 June, suggest that 79% of people worldwide agree, to some extent, that vaccinations are safe.

Europe has some of the lowest

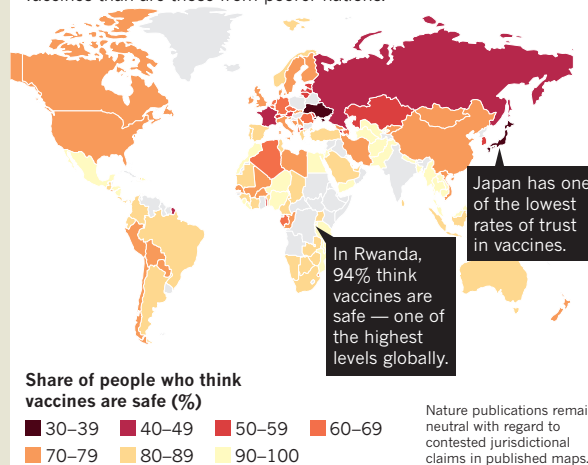
levels of perceived safety, with just half of those in Eastern Europe and 59% in Western Europe strongly or somewhat agreeing with the statement "vaccines are safe". Confidence is particularly low in France and Ukraine.

By contrast, 97% of people in Bangladesh think that vaccines are safe, and that number remains high throughout South Asia. In East Africa, the figure is 92%.

In some regions, greater scientific knowledge was associated with less confidence in vaccines, suggesting that providing information and education might not be enough to combat scepticism.

VACCINE PERCEPTIONS

A global survey of more than 140,000 people funded by the London-based biomedical charity Wellcome has found that people from wealthier countries are more likely to question the safety of vaccines than are those from poorer nations.



NEWS IN FOCUS

ETHICS Africa's science academy tackles unfair research practices **p.284**

TECHNOLOGY Modified PET scanner images whole human body in seconds **p.285**

AGRICULTURE A hungry caterpillar is devastating crops in China **p.286**

GENE EDITING When will the world be ready for CRISPR babies? **p.293**



JOHN WESSELS/AFP/GETTY



Efforts to contain the current Ebola outbreak have been hampered by violence.

PUBLIC HEALTH

WHO resists declaring Ebola emergency

Outbreak worsens in the Democratic Republic of the Congo as virus spreads into Uganda.

BY AMY MAXMEN IN KINSHASA
AND SARA REARDON IN WASHINGTON DC

The Ebola outbreak that has ravaged the northeastern Democratic Republic of the Congo (DRC) — now the second largest such event on record — spread into neighbouring Uganda last week. But for the World Health Organization (WHO), the crisis still does not warrant the highest

level of alarm, said director Tedros Adhanom Ghebreyesus on 14 June.

“Although the spread of Ebola to Uganda is tragic, it is not a surprise,” Ghebreyesus said. “Since the beginning of the outbreak, we have said that the risk of spread across the border is high, and it remains high.”

More than 2,100 people have been infected with

This story was supported by the Pulitzer Center on Crisis Reporting.

the virus since the outbreak began in August 2018, and roughly 1,400 had died as of 12 June, according to the WHO. These figures include the first people to die from Ebola in Uganda during this outbreak: a five-year-old boy and his grandmother, who passed away last week.

The WHO's announcement marks the third time since the outbreak began that the agency has decided against declaring a public-health emergency of international

► concern (PHEIC). The handful of Ebola cases confirmed in Uganda have involved people who travelled from the DRC, and there is no evidence that the virus is being transmitted within Uganda. One criterion that the WHO uses to determine whether an outbreak is a global emergency is whether a disease is spreading in more than one country.

The WHO's Emergency Committee of independent medical experts decided against declaring an emergency in part because doing so could trigger the DRC's neighbours to close their borders. That could halt trade and damage the country's economy, while preventing thousands of people from escaping violence in the northeastern DRC, which is home to dozens of armed groups. At least 50 people

were killed during recent violence in Ituri, one of two provinces in which Ebola is spreading, the DRC government said on 13 June.

"Of course it is an emergency of international concern, translated literally," says Chikwe Ihekweazu, director-general of the Nigeria Centre for Disease Control in Abuja, who advised the deliberations. "But making this additional declaration — to be honest, I don't see the extra benefit it would bring."

Other experts argue that an emergency declaration would help to bring the current Ebola outbreak under control. "I'm baffled," says Lawrence Gostin, a health-law and policy specialist at Georgetown University in Washington DC. "The WHO was roundly criticized for delaying for six months its

declaration of an emergency in West Africa, and now it's repeating history."

He had hoped that an emergency declaration for the DRC outbreak would prompt an outpouring of financial help and other assistance from governments and aid organizations — similar to that seen during the 2014–16 Ebola outbreak in West Africa. The WHO has said that it needs another US\$54 million to support its Ebola response until the end of July.

But Ghebreyesus says that the declaration of a public-health emergency should not be viewed as a fundraising exercise. "You don't wait to patch the roof until after the storm comes," he told *Nature*. "Using a PHEIC to mobilize resources is dangerous because, by then, it's too late." ■

ETHICS

African academy leads push for ethical data use

The goal is to create the continent's first cross-disciplinary guidelines for sharing data.

BY LINDA NORDLING

The African Academy of Sciences (AAS) has started work on the continent's first cross-disciplinary guidelines on how to collect, store and share research data and specimens in ways that protect study participants from exploitation and benefit African citizens.

Members of the AAS Data and Biospecimen Governance Committee, who met for the first time on 10 and 11 June in Nairobi, Kenya, hope

to address these issues, which pose persistent challenges for African nations and researchers. The committee's guidance won't have legal authority — rather, the goal is to provide a resource for governments creating their own policies and to guide researchers, according to committee members.

The AAS holds significant political clout on the continent, says John Mugabe, a science-policy specialist at the University of Pretoria in South Africa. He says that some of the AAS

fellows are senior government officials in their countries or members of the countries' legislative assemblies, and that the academy holds observer status in the African Union.

The AAS committee includes about a dozen African bioethicists, data specialists and legal experts. And it plans to gather input from other disciplines and groups, including patient organizations and community advocates. "The composition of this is basically African, to hear the African voice," says Jenniffer Mabuka-Maroa, an AAS research programme manager based in Nairobi and the convener of the committee meeting.

ETHICS CODES

Other groups and communities in Africa have produced data-sharing guidance or ethics codes. But the AAS committee's work is the first attempt at multidisciplinary guidelines for all of Africa.

At the meeting last week, the group reviewed common challenges for collecting and handling research data in Africa, including sensitive and potentially lucrative information gathered for medical research and bioprospecting — scanning natural resources such as plants and animals for compounds that can be turned into drugs or other commercial products.

Data sharing, particularly in genomics and biodiversity, is a hot topic across Africa. In South Africa, an information-protection bill due to come into effect next year could



African science academy is pushing for better data-sharing guidance for fields including health research.

PIUS UTO MI EKPE/AP/GETTY

block research for which scientists collect data or biological samples such as blood and tissue under 'broad consent', whereby study participants agree that researchers can keep and analyse their samples for unspecified purposes. Some local scientists are lobbying the government for an exemption, saying that restrictions could hamper work on diseases such as HIV and tuberculosis.

Procedures during disease outbreaks are also controversial. Foreign medical workers and researchers who came to West Africa during the 2014–16 Ebola outbreak exported blood samples from some of the affected nations — sometimes without the donors' consent — to their home countries for quicker analysis. Despite attempts to recover these specimens, or to find out what happened to them, many samples remain in foreign laboratories, and it's unclear whether Africans will share in any benefits from the research, such as new treatments.

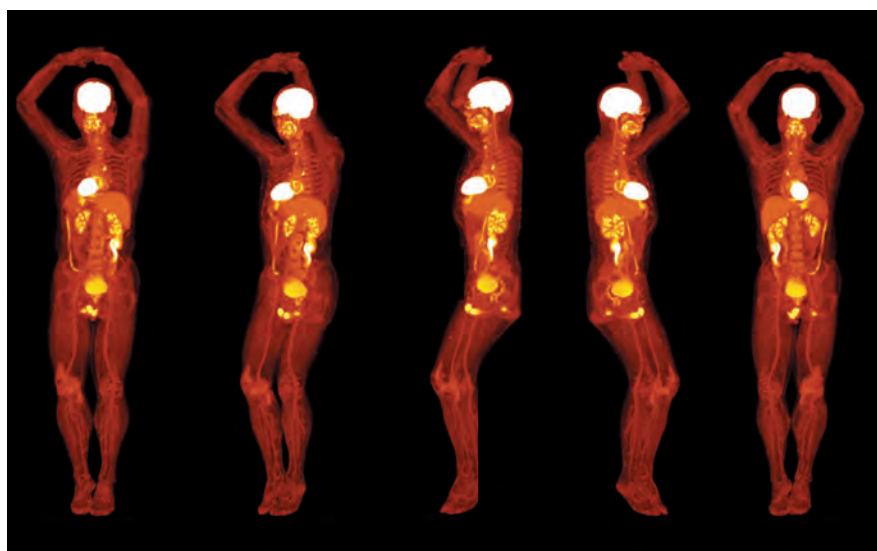
These issues are among many that the AAS committee might consider while drawing up its guidelines.

Even when data-sharing policies are negotiated as part of international partnerships, African scientists might feel unable to push back against the wishes of powerful donors or research partners. African scientists might not question unfair policies for fear of harming their funding chances — and instead simply accept the grants and their inequitable conditions, says an African researcher who didn't want to be named because of the sensitivities of the situation. This is especially problematic for younger researchers, the scientist says.

Clear national or institutional data-sharing policies could help scientists who find themselves in such a position, the AAS committee said. "By having these frameworks, researchers will become less vulnerable and be better able to engage with their donors to achieve mutual respect and benefit."

Whatever way the AAS committee guidelines address the practical challenges of data sharing, it's important that they also uphold African values, says Collet Dandara, a geneticist at the University of Cape Town in South Africa. Communities that take a group-centred approach to participation in research projects, such as the San people in southern Africa, can run into problems with the individualized approach to data ownership and benefit sharing that is common in Western science. Deciding how to proceed would involve working closely with communities that have some claim to the information being gathered, Dandara says, including asking members how they think the work should be done.

Some people say the communities wouldn't understand the research, Dandara says. "But if they don't understand, why are we researching them? Maybe it's us who don't understand." ■



Images from the device can be used to study how drugs or infections move through the human body.

TECHNOLOGY

Whole-body scans made in seconds

A modified PET scanner requires less radiation exposure, vastly broadening its applications.

BY SARA REARDON

A medical imaging device that can create 3D renderings of the entire human body in as little as 20 seconds could soon be used for a wide variety of research and clinical applications.

The modified positron emission tomography (PET) scanner is faster than conventional PET scans — which can take an average of 20 minutes — and requires less radiation exposure for the person being imaged. Researchers presented video taken by the device at the US National Institutes of Health's High-Risk, High-Reward Research Symposium in Bethesda, Maryland, earlier this month.

The machine could be especially helpful for imaging children, who tend to wiggle around inside a scanner and ruin the measurements, as well as for studies of how drugs move through the body, says Sanjay Jain, a paediatrician and infectious-disease physician at Johns Hopkins University in Baltimore, Maryland.

Standard PET scanners detect γ -rays from radioactive tracers that doctors inject into the person being imaged. The person's cells take up the molecules, which release γ -rays, and break them down. A ring-shaped detector positioned around the person measures the

angle and speed of the rays and reconstructs their origin, creating a 3D map of the cells that are metabolizing the molecule.

The ring is just 25 centimetres thick, however, so physicians can image only a small portion of the body at a time. Capturing larger areas requires them to dose the person with more of the radioactive molecule — it decays quickly, which means the signal fades fast — and move them back and forth through the ring.

"The whole-body machine is another quantum jump in medical imaging."

Biomedical engineer Ramsey Badawi and his colleagues at the University of California, Davis, solved this problem by connecting eight PET scanner rings into a 2-metre-long tube that can image the entire body at once. It creates a rendering in 1/40 of the time of a conventional scanner, using 1/40 of the radiation dose and so reducing the radiation risk. The researchers can also leave someone in the scanner for longer periods and take motion-capture images to follow a radioactive tracer through their body.

The US Food and Drug Administration approved the modified scanner for use in the United States last December, and Badawi plans to scan the first patients ►

► with it in California next month.

“The whole-body machine is another quantum jump in medical imaging,” says Abass Alavi, a radiologist at the University of Pennsylvania in Philadelphia. He is collaborating with Badawi to use the modified PET scanner to study atherosclerosis, a condition in which plaque builds up in a person’s arteries.

Eventually, Alavi says, physicians might be able to use the device to see whether certain drugs help to treat the artery-clogging disease.

Conventional PET scanners aren’t usually used for this purpose because of the cost and the radiation exposure to the person, says Badawi.

Jain is hoping to use the device to test a

radioactive sugar tracer he’s developed that’s ingested by bacterial but not mammalian cells. Injecting the tracer into people suspected of having a bacterial infection could highlight where in the body the bacteria are concentrated. Jain’s lab is also developing tracers that could distinguish between types of bacterium. ■

AGRICULTURE

China seeks predator to stop voracious caterpillar

Scientists scramble to find ways to halt the fall armyworm’s march across the country.

BY ANDREW SILVER

A hungry caterpillar that ravages crops is advancing across China and threatening the nation’s vast supply of maize. Scientists are investigating ways to minimize the damage caused by the invasive fall armyworm — which was first detected in China in January — including experimenting with native predators that could keep the pest in check. Some researchers say that the insect’s spread might have been slowed if the country grew genetically modified food crops.

The fall armyworm (*Spodoptera frugiperda*), a native of Central and South America, has spread around the world in the past few years, causing devastation of crops in parts of Africa and southern Asia. Since its arrival in China, it has been found in 18 provinces, regions and municipalities, according to China’s ministry of agriculture.

So far, damage caused by the caterpillar — mostly to maize (corn), but also to other crops such as sugar cane — in China is considered manageable. But Hu Gao, an entomologist at Nanjing Agricultural University who is monitoring the insect’s spread, says researchers and farmers fear what will happen when the pest arrives, probably later this month, at the North China Plain. China is the world’s second-largest producer of maize, and the northern plain produces almost 30% of the country’s crop.

HIGH COST

Recent outbreaks of fall armyworm in Africa and southern Asia have resulted in maize yield losses as high as 50%. In Africa, where the pest arrived in 2016, it costs 12 major maize-growing countries a total of between US\$1 billion and \$4 billion in lost crops a year (X.-J. Li *et al.* Preprint at bioRxiv <http://doi.org/c7dc>; 2019). China is also still battling an epidemic



The invasive fall armyworm has laid waste to crop plants around the world.

of a highly contagious virus affecting pigs, African swine fever, which has led to the culling of more than one million of the animals.

“The spread of fall armyworm in China will have a significant impact, along with the spread of swine fever, on Chinese consumers,” says Cong Cao, a researcher studying innovation at the University of Nottingham Ningbo China. The rising price of food will put tremendous pressure on the government to control the pest, says Cao.

Hu says that plant-protection centres in provinces and cities are focused on monitoring and controlling the fall armyworm’s spread. The adult moths, which are responsible for the pest’s spread, can travel hundreds of kilometres over successive nights. Control measures include traps and pesticides.

Scientists, meanwhile, are working on other strategies. Hu says researchers at his and other Chinese universities are studying chemicals that could be used to attract the insect to traps, and native insects that could be deployed as a means of biological control.

A report by the US Department of Agriculture, released in May, on the fall armyworm’s spread in China said that the insect has no natural predators in the country, but Hu disputes that conclusion.

China’s parasitic Braconid wasps already kill other species in the *Spodoptera* genus to which the fall armyworm belongs, including the cotton leafworm (*Spodoptera litura*) and the beet armyworm (*Spodoptera exigua*), so Hu thinks the wasp could also target the fall armyworm caterpillar. In Africa, some parasitoids

WALDO SWIEGERS/BLOOMBERG/GETTY

— whose young feed on and eventually kill their hosts — that target African cotton leafworm (*Spodoptera littoralis*) have already switched to feasting on the fall armyworm.

During recent field trials in Yunnan Province, where the pest was first identified, researchers based at the Institute of Plant Protection (IPP), part of the Chinese Academy of Agricultural Sciences in Beijing, have also found that the stink bug *Arma chinensis* kills the caterpillar.

PEST CONTROL

There could be many natural parasites or predators that target the pest, says Zhong Guohua, a researcher at the South China Agricultural University in Guangzhou who is working on controlling the fall armyworm, but whether

they can ultimately be used for control it is difficult to predict. Finding out would require repeated testing to ensure that the predator is effective across large areas, and can be bred in large enough numbers, says Zhong.

In some countries, such as Brazil, the pest has been managed by growing transgenic food crops that contain genes from the bacterium *Bacillus thuringiensis* (Bt). The genes offer crops resistance to some pests, including the fall armyworm.

But Bt food crops have not been approved for commercial use in China, in part because

“The spread of fall armyworm in China will have a significant impact on consumers.”

of strong public opposition to genetically modified food, says Du Li, a specialist in biotechnology law at the University of Macau.

The growth of Bt maize across a large area of China would definitely have helped to control the pest, says Li Yunhe, a biotechnology researcher at the IPP.

But Hu says that it's not clear whether the crop can keep the pest at bay in the long term. In countries such as the United States, the insect has developed resistance to Bt crops, he notes.

Hu says that eradication in China is now unlikely, and that farmers will have to learn to manage the pest. Other major crop-producing countries are also in the insect's path — researchers predict that it will probably enter Japan and South Korea between now and next month. ■

RESEARCH MISCONDUCT

What universities can learn from epic case of research fraud

Analysis of misconduct investigations suggests institutional probes aren't rigorous enough.

BY HOLLY ELSE

By day, Andrew Grey studies bone health. But over the past few years, he's developed another speciality: the case of one of science's most prolific fraudsters.

From 1996 to 2013, Yoshihiro Sato, a Japanese bone-health researcher, plagiarized work, fabricated data and forged authorships — prompting the retraction of more than 60 studies from the scholarly literature so far. Grey and colleagues at the University of Auckland in New Zealand and the University of Aberdeen, UK, are among the researchers who have raised concerns about Sato's work over the past decade or so, and they have studied the case in detail — in particular, how universities involved in the research investigated concerns about his work and allegations of misconduct.

At the World Conference on Research Integrity in Hong Kong from 2 to 5 June, Grey's team described its years-long efforts to clean up Sato's literature, and presented its analysis of the inquiries conducted by four universities in Japan and the United States ensnared in the scandal. The team published its analysis of three investigations in February (*A. Grey et al. Res. Integr. Peer Rev.* 4, 3; 2019). Grey says the findings support a growing view among some in the academic community: that university investigations into research misconduct are often inadequate, opaque and poorly conducted. The team says that the results challenge the idea that institutions can police themselves on research

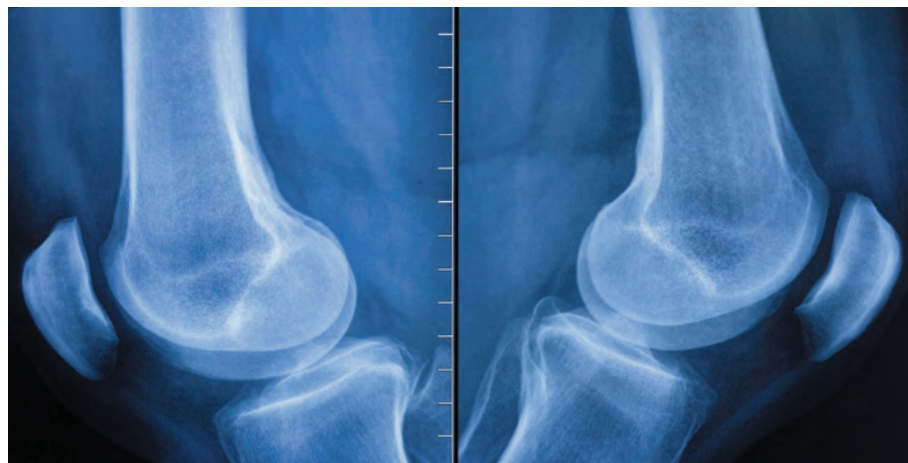
integrity, and proposes that there should be independent organizations to evaluate allegations of research fraud.

The analysis is one of just a few to look closely at research-misconduct investigations, and the first to use a systematic approach to rate them, says C. K. Gunsalus, a specialist in research integrity at the University of Illinois at Urbana-Champaign, who was not part of the analysis. Too many research-misconduct investigations turn out to be inadequate or flawed, says Gunsalus. She had a hand in creating a 26-point checklist that university officials can use to guide probes into research misconduct; Grey's team used this to rate the investigations.

The checklist questions an investigation's scope, reliability and impact — for instance, whether the investigating committee included external members and whether evidence could have been tampered with. Two members of Grey's team independently assessed each investigation report using the checklist. “Overall, each report was considered unacceptable by both assessors,” say Grey and colleagues.

ALARM BELLS

Sato, who died in 2016, studied and ran clinical trials of drugs and supplements that might help to prevent bone fractures. Researchers in the field began raising concerns about his work ▶



The field of bone health was hit by a sprawling case of research misconduct that affected tens of studies.

► in the mid-2000s, when some questioned the speed with which Sato had recruited and assessed participants for some of his studies. He later apologized for not disclosing all the hospitals from which he had recruited participants, and admitted that there was a mistake in one paper. But more researchers started flagging irregularities in his papers to journals, and, in 2016, Grey and his colleagues published an analysis that raised concerns about 33 of Sato's studies (M. J. Bolland *et al. Neurology* 87, 2391–2402; 2016). Sato admitted that three of these studies were fraudulent, asked for them to be retracted and cleared his co-authors of any wrongdoing. Twenty-seven of the studies have now been retracted.

In 2017, Grey's team also flagged concerns about hundreds of Sato's papers to four institutions that employed co-authors of these studies: Kurume, Hirosaki and Keio universities in Japan, and New York University's Winthrop Hospital. Sato had been a researcher at Kurume University. Two institutions had already launched investigations into some of the work when Grey contacted them, and the others began investigations.

The researchers asked the institutions for the reports of their investigations, to understand how they had responded to the allegations. None of the reports revealed exactly who or which papers had been investigated.

One found that an unnamed researcher had committed misconduct; two reports recommended that papers be retracted.

Grey's team rated each report as inadequate overall. The researchers also suggest that the investigations focused too much on determining whether research misconduct had occurred, rather than on understanding the validity of the research in question and correcting or retracting unreliable articles. Grey and his colleagues argue that protecting the integrity of the literature should be the priority of any investigation — because integrity can be compromised without evidence of misconduct.

FURTHER REVIEW

Gunsalus agrees that the Sato case highlights some of the problems with misconduct investigations, and says that if shortcomings emerge, further reviews might be needed. She suggests that institutional panels should include external members, and that officials should use a standardized checklist to strengthen their processes. “There should be some way for journals, funders, patients and others to be assured of the credibility and thoroughness of

university reviews,” she says.

Grey's findings also suggest that institutions in Japan — which has seen several high-profile research-misconduct cases in recent decades — should review their processes for investigating misconduct, says Alan Price, a research-misconduct consultant in Texas.

The universities did not respond directly to criticisms of the investigations, which *Nature* flagged to them, but offered further details about their inquiries and the outcomes. Winthrop Hospital said that it spent more than a year investigating the concerns, including digging up receipts for lab equipment, but found no misconduct. Keio University said that its investigation included external experts and statistical analysis of data; it found no research misconduct, but some errors in methods and typos in studies.

Kurume University asked a committee of statisticians and medical researchers to investigate 39 papers authored by Sato, and found some data falsification and inappropriate authorships. It said that it cannot conclude whether fraud was involved in another 32 papers, because Sato is dead and records for these experiments no longer exist. Hirosaki University — whose 2017 investigation found research irregularities in 14 research papers, 7 of which had already been retracted — did not respond to *Nature's* request for comment. ■

nature research ACADEMIES



Training workshops for researchers

A series of workshops to support researchers, covering topics such as getting published, journal editing, clinical research methodology, and applying for research positions.

Visit partnerships.nature.com/academies to host an academy at your institution.



The secret social lives of viruses

Long thought the lone wolves of the microbial world, viruses are turning out to be surprisingly sociable. Listening in could be the key to fighting infection.

BY ELIE DOLGIN

Geneticist Rotem Sorek could see that his bacteria were sick — so far, so good. He had deliberately infected them with a virus to test whether each ailing microbe soldiered on alone or communicated with its allies to fight the attack.

But when he and his team at the Weizmann Institute of Science in Rehovot, Israel, looked into the contents of their flasks, they saw something completely unexpected: the bacteria were silent, and it was the viruses that were chattering away, passing notes to each other in a molecular language only they could understand. They were deciding together when to lie low in the host cell and when to

replicate and burst out, in search of new victims.

It was an accidental discovery that would fundamentally change scientists' understanding of how viruses behave.

Viruses that infect bacteria — spiky lollipop-like creatures known as bacteriophages (or phages) — have surveillance mechanisms that bring them intel on whether to stay dormant or attack, depending on the availability of fresh victims. But researchers long thought these processes were passive; the phages seemed to just sit back and listen in, waiting for bacterial distress signals to reach fever pitch before taking action.

Sorek and his colleagues had found phages actively

ILLUSTRATION BY KAROL BANACH

discussing their choices. They realized that as a phage infects a cell, it releases a tiny protein — a peptide just six amino acids long — that serves as a message to its brethren: “I’ve taken a victim”. As the phages infect more cells, the message gets louder, signalling that uninfected hosts are becoming scarce. Phages then put a halt to lysis — the process of replicating and breaking out of their hosts — instead staying hidden in a sluggish state called lysogeny¹.

The viruses, it turns out, did not depend on bacterial cues to make their decisions. They controlled their own destiny. “This finding was a big, important, revolutionary concept in virology,” says Wei Cheng, a structural microbiologist at Sichuan University in Chengdu, China.

Sorek named this viral peptide ‘arbitrium’, after the Latin word for decision. It seemed to work much like the communication system used by bacteria — quorum sensing — to share information about cell density and adjust the population accordingly. Yet it was the first time anyone had demonstrated molecular messaging of this kind in viruses. And it fitted into an emerging picture of viruses as much more sophisticated social agents than scientists had given them credit for.

Virologists have long studied their subjects in isolation, targeting cells with just a single viral particle. But it’s become increasingly clear that many viruses cooperate, teaming up to co-infect hosts and break down antiviral immune defences.

The implication is that researchers might have been going about their experiments all wrong. “It has shaken one of the pillars of virology,” says Sam Díaz-Muñoz, an evolutionary biologist at the University of California, Davis.

Learning the language behind these viral interactions could inform the design of new treatments for cancer and nasty superinfections. The social predilections of viruses even help to explain how they evade the bacterial immune system known as CRISPR. “Conceptually, it’s really powerful,” Díaz-Muñoz says.

SOCIAL STUDIES

Scientists first spied viruses mingling in the 1940s, when separate experiments by biophysicist Max Delbrück and bacteriologist Alfred Hershey showed that two viral particles could simultaneously invade the same cell and swap genes. But according to Dale Kaiser, a molecular geneticist at Stanford University in California and a protégé of Delbrück’s, these early observations were only really interesting to scientists as an experimental method — they allowed researchers to create a cross between two viral strains. The relevance to basic biology was missed.

It wasn’t until 1999 that anyone took any notice of what cooperation achieved for the viruses themselves. That year, evolutionary biologists Paul Turner, now at Yale University in New Haven, Connecticut, and Lin Chao, now at the University of California, San Diego, showed that phages play their own version of the prisoner’s dilemma strategy game, working in partnership under certain circumstances and acting in their own self-interests in others².

Other examples of beneficial viral interactions followed, including ones that involved the pathogens responsible for diseases such as hepatitis, polio, measles and influenza. They often took place between different viral strains that had a shared interest in boosting their own reproductive chances. But the molecular basis of those cooperative traits — the method of communication — had largely remained elusive. And as Rafael Sanjuán, an evolutionary geneticist at the University of Valencia in Spain, points out: “The ‘how’ is really important here.”

That’s why the arbitrium discovery was such a big step forward for the field.

Almost immediately after Sorek first described the phenomenon, in 2017, four independent groups — including Cheng’s and one led by structural biologist Alberto Marina at the Biomedical Institute of Valencia in Spain — set to work trying to reveal the molecular basis by which arbitrium peptides are made, sensed and acted on by phages.

Those technical details, reported in five papers^{3–7} over the past nine months, helped to explain exactly how the short peptides Sorek discovered influence viral decision-making. For Marina, however, this is just the start of the story: he suspects that the communication system probably serves many more functions.

Marina’s suspicion rests on a finding in one of those papers⁶. Working with José Penadés, a microbiologist at the University of Glasgow, UK, Marina showed that the receptor for arbitrium in the phage can interface not only with genes in the bacterium that help the virus to reproduce, but also with other, unrelated stretches of DNA. That means that its activity might not be limited to the virus’ stay-or-go decision. The researchers are now exploring whether the phage’s peptide language alters the activity of key genes in its victim, too. “If true,” Marina says, “this would make the picture much bigger and more exciting.”

Expanding on his own initial discovery, Sorek has found arbitrium peptides popping up everywhere. His team has now found at least 15 different types of phage, all of which can infect soil microbes and use some sort of short peptide to communicate⁸. Notably, says Sorek, “each phage seems to speak in a different language and only understands its own one”. The viral chit-chat thus seems to have evolved to allow communication only between close relatives.

Phages might speak only to their own kind, but they can also listen in on other languages. Molecular biologist Bonnie Bassler and her graduate student Justin Silpe have found that viruses can use quorum-sensing chemicals released by bacteria to determine when best to start multiplying — and murdering⁹. “The phages are eavesdropping, and they’re hijacking host information for their own purposes — in this case, to kill the host,” Bassler explains.

This molecular snooping occurs naturally in phages that infect the bacterium responsible for cholera, *Vibrio cholerae*. But in their lab at Princeton University in New Jersey, Bassler and Silpe have engineered ‘spy’ phages that can sense signals unique to other microbes, including *Escherichia coli* and *Salmonella typhimurium*, and obliterate them. The viruses in effect became programmable assassins that could be made to kill off any bacterium — at will and on demand.

FOR THE GREATER GOOD

Some viral cooperation seems to verge on altruism. Two independent groups reported last year that some phages act selflessly to overcome the viral countermeasures of *Pseudomonas bacteria*^{10,11}.

The teams — one led by phage biologist Joe Bondy-Denomy at the University of California, San Francisco, the other by CRISPR expert Edze Westra and virologist Stineke van Houte at the University of Exeter, UK — watched as viruses bombarded bacteria with specialized proteins designed to break down the cells’ CRISPR-based immune defences. The first wave of viruses attacked the cells, killing themselves but also weakening the bacteria. The initial bombardment paved the way for others to conquer the microbial foe. “Those phages had to be there, and to die, and produce anti-CRISPRs before another phage could come along and succeed,” says Bondy-Denomy.

In follow-up work, Westra and his postdoc Anne Chevallereau demonstrated how phages lacking these anti-CRISPR proteins can exploit the cooperative offerings of others that do¹². To Westra, that shows the potentially far-reaching consequences of altruistic behaviours among viruses. “There are a lot of emergent properties at the population level,” he says. “It’s very important to keep the ecology of these phages in mind.”

These examples of communication and cooperation in phages are probably just the tip of the social spear, says Lanying Zeng, a biophysicist at Texas A&M University’s Center for Phage Technology in College Station. “This is a whole unexplored area.” And the same goes for viruses that infect other cell types — including animal and human cells — which employ some social tricks of their own.

Take vesicular stomatitis virus (VSV), which mainly infects farm

“It has shaken one of the pillars of virology.”

SPL animals, but can cause a flu-like illness in humans, too. Particles of this viral pathogen suppress host immunity at a personal cost but at a benefit to the group, as Sanjuán and his colleagues have shown¹³. No one is sure yet how this cooperative evasion is happening, but the work highlights how crucial altruism can be for the success of VSV. That could help scientists to beat the virus in farm animals, and optimize it for use in vaccines and therapeutics.

Other instances of collective action are widespread among disease-causing viruses. In poliovirus, for example, multiple genetically distinct viral strains can clump together to swap gene products and enhance their human-cell-killing potential¹⁴. And two strains of influenza — one that excels at cell entry, the other at cell exit — grow better when maintained in cell culture together than when kept apart¹⁵.

But in a real-world setting, in nasal swabs from people with influenza, the two viral strains didn't seem to coexist¹⁶. Jesse Bloom at the Fred Hutchinson Cancer Research Center in Seattle, Washington, who led the research, thinks that has to do with some peculiarities of the flu virus' life — its population size swings so wildly that cooperative particles have a slim chance of sticking together. For viruses that don't undergo those kinds of transmission bottlenecks, "cooperation might be more likely to be maintained in real-world settings", he says.

That's exactly what microscopist Nihal Altan-Bonnet found when she studied rotavirus transmission between mouse pups. Rotavirus particles can travel together between cells in bubble-like vesicles, sharing resources and hiding from the host's immune system. And, Altan-Bonnet and her colleagues have shown, the particles become more infectious to mice when they are inside these cooperative clusters than when going it alone¹⁷.

Many other pathogenic viruses — including those responsible for Zika, hepatitis, chickenpox, norovirus and the common cold — are now known to transmit themselves through these vesicles, too.

"These viruses are very sneaky," says Altan-Bonnet, who heads the Laboratory of Host-Pathogen Dynamics at the US National Heart, Lung, and Blood Institute in Bethesda, Maryland. "And we have to think of strategies that disrupt this cooperativity and clustering of viruses."

That is, unless the destructive power of viruses could be used for good. Several groups are testing phages as a treatment for bacterial infections — and knowing more about how they converse with each other could help to refine such therapies, which have a long history in medicine but are only just starting to be manipulated for therapeutic gain.

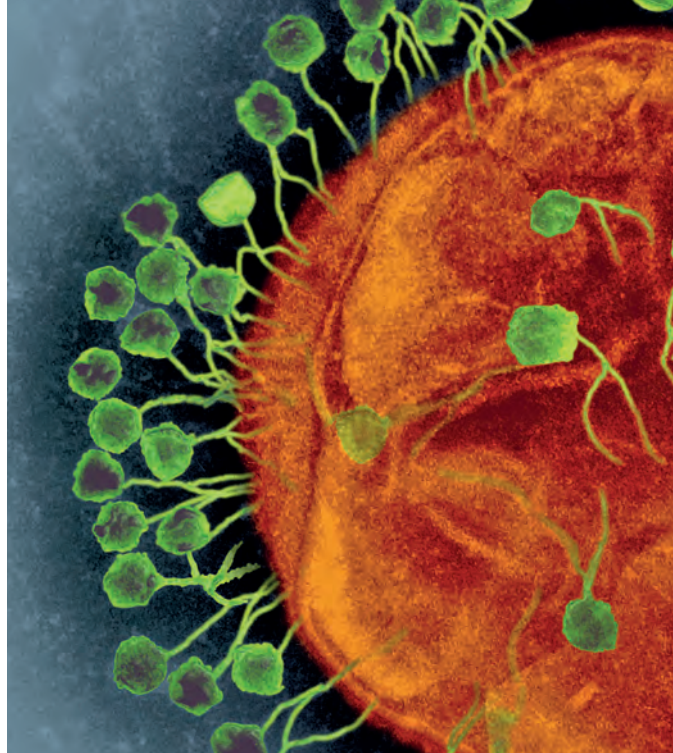
ENGAGE THE PHAGE

Last month, for instance, researchers described the first successful clinical use of genetically engineered phages to tackle a drug-resistant bacterial infection¹⁸. For infections such as this, of course, the ideal solution is to use the virus to annihilate the bacteria entirely. But for conditions that are marked by a microbial imbalance, such as acne, some types of cancer and inflammatory bowel disease, it might be better to deploy a phage that can help to restore the balance without an all-out assault.

And for those more subtle applications, knowing exactly how viruses communicate "could be really useful for helping us to engineer phages that could be used for treating disease", says Karen Maxwell, a phage biologist at the University of Toronto in Canada. Tapping into the arbitrium system could thus lead to more tractable, or even reversible, treatments.

Learning to speak virus could provide a different kind of therapeutic benefit, too. "This could be an addition to the synthetic-biology toolkit to help fine-tune engineered bacterial gene expression," says Christopher Alteri, a microbiologist at the University of Michigan in Dearborn.

Sorek, for example, has taken the arbitrium peptides out of their natural habitat in the phage and plugged them into other organisms, where they act as dimmer switches that dial up or dampen gene activity. In unpublished work, he and his graduate student Zohar Erez inserted the arbitrium machinery into the bacterium *Bacillus subtilis*, allowing them to manipulate several of its genes at will. The engineered microbes could one day be used, for instance, to deliver medicines in precise doses or to specific locations.



Viruses known as phages (green) can better infect cells like this bacterium (orange) when they cooperate and communicate.

What's more, notes Sorek, if arbitrium-like systems turn out to be conserved in human viruses — pathogens such as HIV and herpes simplex virus that, like phages, spend portions of their lives hiding out in cells — then any communication molecule that prompts viral dormancy "immediately becomes a drug".

Every scientific project that persists gets an 'ology', and the study of sociable viruses is no different. Two years ago, Díaz-Muñoz, Sanjuán and evolutionary biologist Stu West from the University of Oxford, UK, coined¹⁹ a new term — sociovirology — to provide a framework for their line of research. The American Society for Microbiology will host the first-ever workshop dedicated to the topic at its annual meeting this month in San Francisco. "It's an idea whose time has come," Díaz-Muñoz says.

In sociovirology, he sees many parallels with the gradual acceptance of similar group behaviours among bacteria in years past: it wasn't until researchers pinpointed the chemicals involved in quorum sensing and put a name to the process that most microbiologists paid the phenomenon any attention.

"It isn't in the consciousness," Díaz-Muñoz says. But as with all things social and viral, the message is spreading. ■

Elie Dolgin is a science journalist in Somerville, Massachusetts.

1. Erez, Z. *et al. Nature* **541**, 488–493 (2017).
2. Turner, P. E. & Chao, L. *Nature* **398**, 441–443 (1999).
3. Wang, Q. *et al. Nature Microbiol.* **3**, 1266–1273 (2018).
4. Dou, C. *et al. Nature Microbiol.* **3**, 1285–1294 (2018).
5. Zhen, X. *et al. Protein Cell* **10**, 131–136 (2019).
6. Gallego Del Sol, F., Penadés, J. R. & Marina, A. *Mol. Cell* **74**, 59–72.e3 (2019).
7. Guan, Z. *et al. Cell Discov.* **5**, 29 (2019).
8. Stokar-Avihail, A., Tal, N., Erez, Z., Lopatina, A. & Sorek, R. *Cell Host Microbe* **25**, 746–755.e5 (2019).
9. Silpe, J. E. & Bassler, B. L. *Cell* **176**, 268–280.e13 (2019).
10. Landsberger M. *et al. Cell* **174**, 908–916.e12 (2018).
11. Borges, A. L. *et al. Cell* **174**, 917–925.e10 (2018).
12. Chevallereau, A. *et al.* preprint at bioRxiv <https://doi.org/10.1101/574418> (2019).
13. Domingo-Calap, P., Segredo-Otero, E., Durán-Moreno, M. & Sanjuán, R. *Nature Microbiol.* **4**, 1006–1013 (2019).
14. Aguilera, E. R., Erickson, A. K., Jesudhasan, P. R., Robinson, C. M. & Pfeiffer, J. K. *mBio* **8**, e02020-16 (2017).
15. Xue, K. S., Hooper, K. A., Ollodart, A. R., Dingens, A. S. & Bloom, J. D. *eLife* **5**, e13974 (2016).
16. Xue, K. S., Greninger, A. L., Pérez-Osorio, A. & Bloom, J. D. *mSphere* **3**, e00552-17 (2018).
17. Santiana, M. *et al. Cell Host Microbe* **24**, 208–220.e8 (2018).
18. Dedrick, R. M. *et al. Nature Med.* **25**, 730–733 (2019).
19. Díaz-Muñoz, S. L., Sanjuán, R. & West, S. *Cell Host Microbe* **22**, 437–441 (2017).



Jeff Carroll was concerned about passing Huntington's disease on to his children.

CRISPR BABIES

When will the world be ready?

Scientists say efforts to make heritable changes to the human genome are premature and fraught with uncertainty. Here's what it could take to make the technique safe and acceptable.

BY HEIDI LEDFORD

TAEHOON KIM FOR NATURE

Jeff Carroll had been married for six months when he and his wife decided not to have children. Carroll, 25 years old and a former corporal in the US Army, had just found out that he had the mutation that causes Huntington's disease, a genetic disorder that ravages the brain and nervous system and invariably ends in an early death. He had learnt that his mother had the disease about four years earlier, and now he knew that he was all but certain to develop it, too.

Faced with a 50% chance of passing on the same grim fate to their children, the couple decided that kids were out of the question. "We just kind of shut that down," says Carroll.

But he had begun studying biology in the army in the hope of learning more about the disease. He found out about a process called pre-implantation genetic diagnosis or PGD. By conceiving through *in vitro* fertilization (IVF) and screening the embryos, Carroll and his wife could all but eliminate the chance of passing on the mutation. They decided to give it a shot, and had twins free of the Huntington's mutation in 2006.

Now Carroll is a researcher at Western Washington University in Bellingham, where he uses another technique that might help couples in his position: CRISPR gene editing. He has been using the powerful tool to tweak expression of the gene responsible for Huntington's disease in mouse cells. Because it is caused by a single gene and is so devastating, Huntington's is sometimes held up as an example of a condition in which

gene editing a human embryo — controversial because it would cause changes that would be inherited by future generations — could be really powerful. But the prospect of using CRISPR to alter the gene in human embryos still worries Carroll. "That's a big red line," he says. "I get that people want to go over it — I do, too. But we have to be super humble about this stuff." There could be many unintended consequences, both for the health of individuals and for society. It would take decades of research, he says, before the technology could be used safely.

Public opinion on gene editing to prevent disease is largely positive. But Carroll's reticence is common among scientists. When news broke last year that a Chinese biophysicist had used genome editing in an attempt to make children more resistant to HIV, many scientists were quick to condemn the move as premature and irresponsible.

Several researchers and scientific societies have since called for a moratorium on heritable genome editing in humans. But such a moratorium raises an important question, says embryologist Tony Perry of the University of Bath, UK. "When would it stop?" he asks. "What conditions would you need to meet?"

Nature asked researchers and other stakeholders what hurdles remain before heritable gene editing could become acceptable as a clinical tool. Although some scientific challenges are probably surmountable, approval on a grand scale is likely to require changes to how clinical trials are run, as well as a broader consensus about the technology.

Off-target edits: how many 'mistakes' are too many?

Genome editing presents many difficult technical challenges, but the spectre of creating unwanted genetic changes has probably received the most attention, says Martin Pera, a stem-cell researcher at the Jackson Laboratory in Bar Harbor, Maine. And yet, he adds, this challenge might also be the easiest to surmount.

The most popular way to edit genes relies on a system called CRISPR-Cas9. Co-opted from a mechanism that some microbes use to defend themselves against viruses, it uses an enzyme called Cas9 to make cuts to DNA. A scientist can supply a snippet of RNA to guide Cas9 to a specific site in the genome. But Cas9 and enzymes like it have been known to cut DNA at other sites, too, particularly when there are DNA sequences in the genome similar to the target (see 'Off-target effects'). Such 'off target' cuts could result in health problems: a change to a gene that suppresses tumour growth, for example, might lead to cancer.

Researchers have looked to develop alternatives to the Cas9 enzyme, some of which might be less error-prone. They have also engineered

versions of Cas9 that have lower error rates¹.

Error rates vary depending on what site in the genome is targeted. And many of the gene-editing enzymes have been studied only in mice or in human cells grown in culture — not in human embryos. The rate of mistakes could differ between mice and human cells, and between mature cells and embryos.

The number of errors might not need to be zero. A small number of DNA changes occur naturally every time a cell divides. Some say that a few background changes could be acceptable, especially if the technique is being used to prevent or treat a serious disease.

Some researchers already consider the error rate for CRISPR to be sufficiently low, says Perry. "But, and I think it's quite a big 'but', we don't really have a handle on the editing specificity in human oocytes and embryos," he says.

On-target, but wrong: how precise does gene editing need to be?

A bigger problem than off-target effects might be DNA changes that are on-target but unwanted. After Cas9 or a similar enzyme cuts DNA, it is up to the cell to heal the wound. But the cell's repair processes are unpredictable.

One form of repair, called non-homologous end joining, often deletes some DNA letters at the cut site — a process that could be useful if the goal of the edit is to shut down expression of a mutant gene.

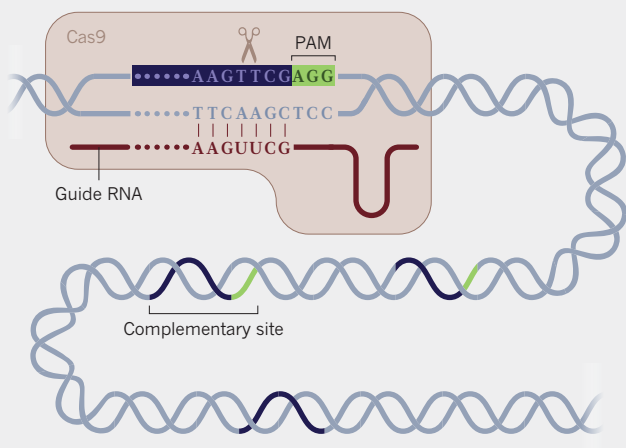
Another form of repair, called homology-directed repair, allows researchers to rewrite a DNA sequence, by supplying a template that gets copied in at the site of the cut. This could be used to correct a disease such as cystic fibrosis, which is generally caused by short deletions in the *CFTR* gene (see 'On-target effects').

Both processes are difficult to control. The deletions caused by non-homologous end joining can vary in size, producing different DNA sequences. Homology-directed repair gives more control over the editing process, but it occurs at a much lower frequency than deletions in many cell types. Research in mice can make CRISPR gene editing seem more precise and efficient than it is now, says Andy Greenfield, a geneticist at the UK Medical Research Council's Harwell Institute near Oxford. Mouse litters are large, and so researchers have a lot of shots at goal to get the right edit — discarding all errors. The same would not be true for human embryos.

It is not yet clear how efficient homology-directed repair would be in humans, or even how it would work. In 2017, one team used

Off-target effects

The Cas9 protein works like a pair of molecular scissors. A guide RNA sequence binds to a complementary DNA sequence that is adjacent to a string of letters known as the proto-spacer adjacent motif (PAM). But there can be many sites in the genome that contain the same or similar sequences, so Cas9 might cut in the wrong places.



CRISPR–Cas9 in human embryos to correct gene variants associated with heart failure². The embryos were never implanted, but the results suggested that the modified cells had used the mother's genome as their template for DNA repair, rather than the DNA template that the researchers had provided. That could be a more reliable way to edit DNA in human embryos. But other researchers have since reported that they have been unable to repeat the results³. “At this point, we don't really understand how embryos deal with DNA repair,” says Jennifer Doudna, a molecular biologist at the University of California, Berkeley. “A lot of work needs to be done in other kinds of embryos, just to understand the fundamentals.”

Researchers are developing ways around the problems associated with DNA repair. Two reports published in June discuss a CRISPR system that can insert DNA into the genome without breaking both strands, thereby bypassing the reliance on DNA repair mechanisms. If the systems hold up to further testing, they could offer researchers greater control over what they edit^{4,5}.

Another approach is to use a technique called base editing. Base editors fuse a disabled Cas9 to an enzyme that can convert one DNA letter to another⁶. The disabled Cas9 directs the base editor to a site in the genome where it chemically changes the DNA directly, rather than by making a break. Studies published in April have shown that some of these base editors are prone to making off-target changes, too^{7,8}, but work is ongoing to try to improve their fidelity.

“Base editing doesn't currently meet our criteria,” says Matthew Porteus, a paediatric haematologist at Stanford University in California. “But one can imagine it getting better and better.”

Wanted, but dangerous: which edits are safe?

Even if the targeting and precision of changes in genome editing were perfect, there would still be a question about what kinds of changes to the human germ line are likely to be safe. In 2017, an international effort spearheaded by the US National Academies of Sciences, Engineering, and Medicine outlined the conditions that should be met before editing a human embryo that is destined for implantation⁹. One of the criteria was that the DNA sequence created by the edit already be common in the population, and carry no known risk of disease.

That requirement alone would put heritable gene editing in people out of reach for the near future, says Porteus. It is not only difficult to predict the precise sequence of an edit, but also hard to know with certainty that a variant will not increase the risk of disease.

Some mutations in a gene called *PCSK*, for example¹⁰, are associated with lower cholesterol levels and therefore a reduced risk of heart disease. The gene is sometimes suggested as a candidate for editing. But only a small number of people have those protective mutations, notes Porteus. The people known to have it are healthy, but researchers don't know how many others might have had the mutation and died.

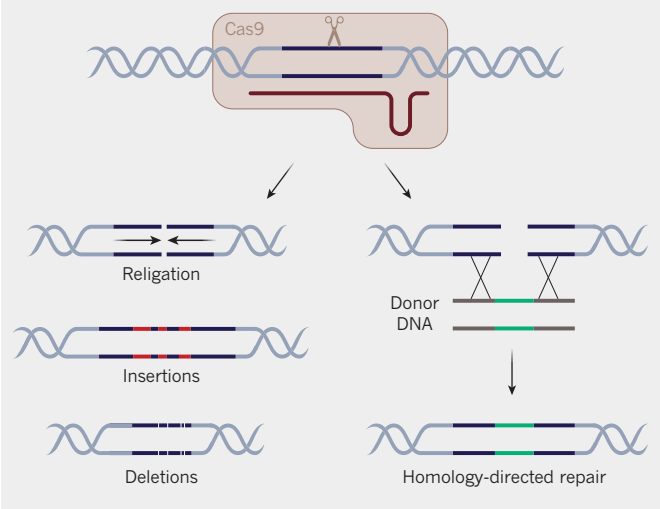
The first known attempt at heritable gene editing in humans was an effort to disable a gene called *CCR5*, which produces an immune-cell receptor that allows HIV to infect humans. Break the gene, and the children should be resistant to the virus, reasoned He Jiankui, then at the Southern University of Science and Technology in Shenzhen, China. He attempted to create a *CCR5* mutation that is found naturally in some people of European descent and is associated with HIV resistance. But a study published this month using data from the UK Biobank found that the deletion might also shorten lifespan⁹.

The effects of some genetic variants can also depend on the environment and on other variants present in the genome. The *CCR5* mutation, for example, is very rare in Chinese populations, raising concerns that the gene could be important for protection against viruses that people would be more likely to encounter in Asia.

That kind of confusion can cause trouble for heritable gene editing, notes Cletus Tandoh Andoh, a bioethicist at the University of Yaoundé in Cameroon. “The majority of studies of genetic association with disease have been performed in Europeans,” he says. To deploy heritable genome editing in Africa, for example, extensive gene and environment studies would first need to be done in African populations, he says.

On-target effects

After Cas9 cuts DNA, the cell tries to repair the damage, but the processes it uses are unpredictable. The fissure can be repaired perfectly (religation) or some letters could be inserted or deleted. Researchers can also introduce donor DNA for the cell to use as a template in homology-directed repair. This process is more precise, but less efficient.



Patchwork babies: how can researchers prevent mosaics?

Sometimes genes differ not only between individuals in a population, but also among the cells of an individual. The advent of cheap and rapid genome sequencing has revealed that this condition, known as mosaicism, is more common than once thought.

Mosaicism might pose problems for gene editing. An embryo tweaked to correct a gene that causes Huntington's disease could contain a mix of corrected and uncorrected cells. How that affects the health of the resulting child would depend on which cells were edited and which were not — something that could be difficult to predict in advance.

Rudolf Jaenisch, a stem-cell scientist at the Whitehead Institute in Cambridge, Massachusetts, doubts that researchers will ever be able to rule out the possibility of mosaicism in an embryo. And methods to assay the DNA sequence in an embryo rely on removing a small number of cells for testing, and then destroying them. Researchers can't test the cells that remain. “Even if you do preimplantation diagnosis,” he says, “it is impossible to decide whether it was a success.”

Some researchers have reported injecting the CRISPR–Cas9 machinery into embryos at very early stages of development², when they are still only a single cell. This technique eliminated mosaicism, the authors said. But it will need to be tested many more times to be sure, says Perry.

And genome editing so early in development creates a new problem: there is no way to distinguish embryos that carry the genetic disease from those that do not at the single-cell stage, cautions Jaenisch. “You will, by definition, manipulate healthy embryos,” says Jaenisch, and so expose them to unnecessary risk (see ‘Mosaicism’).

Would any degree of mosaicism be tolerable? It might depend on the condition being treated, says Krishanu Saha, a bioengineer at the University of Wisconsin–Madison. “If we have 30% of the liver edited and we're trying to treat, let's say, a retinal disease, is that ok?” he says. “In some cases it could be.”

Testing times: how should clinical trials be designed?

With all these technological barriers still to cross, there has been comparatively little discussion of how heritable genome editing would be tested in clinical trials, and what data would be needed before the technique can make that step. The requirements should be high, because the changes could be passed on to future generations, says Guoping Feng, a neuroscientist at the Massachusetts Institute of Technology in Cambridge. “This is not like an ‘I'm going to have a cramp in my

stomach' side effect," he says. "This is permanent."

Some are looking to the example set by the UK Human Fertilisation and Embryology Authority (HFEA), which spent 14 years analysing data from animals and people before it decided to conditionally allow a technique called mitochondrial donation. The technique allows women with disease-causing mutations in the DNA of the cell's power plants — its mitochondria — to use mitochondria from the egg of a healthy donor during IVF. As with gene editing, it could allow parents to avoid passing along dangerous mutations. And there are still questions about the safety of this procedure — some countries, including the United States, do not allow it. Even so, many more data were available about that technique than there are now for CRISPR–Cas9 editing in embryos, says Greenfield, who served on the HFEA panel. (IVF took more than 30 years to move from laboratory testing to a healthy pregnancy.)

Human clinical trials would present a host of fresh challenges. For example, for how long will genome-edited children need to be followed up before the technique can be considered safe? How will researchers track the children of those children to look for transgenerational effects? "It's going to be a mess," says Bryan Cwik, a bioethicist at Portland State University in Oregon.

On 22 May, the US National Academy of Sciences, the US National Academy of Medicine and the UK Royal Society announced a committee to study these aspects of heritable gene editing. The panel aims to publish a report next year. "There is really a need to have a much more in-depth set of criteria in place," says Doudna. "I think we all wish that would have happened faster than it had."

The biggest question: is the world ready?

Despite the sizeable scientific barriers to heritable gene editing, the more difficult issues are likely to be ethical and social. Consultations have been ongoing, and reports and position statements have been pouring in from scientific societies around the world. In March, a panel convened by World Health Organization (WHO) concluded that it would currently be irresponsible to make heritable edits to the genome in humans. Authors writing in *Nature* have called for a global moratorium¹¹, and members of the US National Academy of Sciences, the US National Academy of Medicine and the Royal Society have said that "we must achieve broad societal consensus before making any decisions".

Achieving global consensus is a daunting task and, at present, most of the consultations have been conducted in wealthy, Western countries. Kewal Krishan, an anthropologist at Panjab University in Chandigarh, India, says that there has been little discussion of heritable gene editing in India, for example. And Andoh notes that in some African cultures, the pressure to have children is intense, and women can be ostracized from the community for failing to do so. This could foster demand.

Demand is another question entirely. For now, there is not a huge clamour among people affected by disease, says Sharon Terry, president and chief executive of Genetic Alliance, an advocacy group in Washington DC. Initial enthusiasm has been tempered over time, both as debates advanced and as patient advocates realized that treatments were not imminent, she says. Many families at risk of passing on genetic diseases tell her that, for now, they just want a way to screen their embryos for mutations. But screening is hardly a panacea. It won't work for all couples.

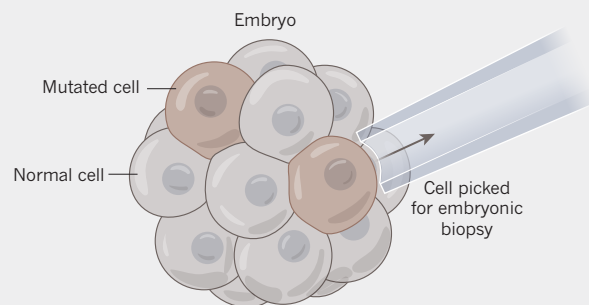
Such decisions are intensely personal, says Andrew Imparato, executive director of the Association of University Centers on Disabilities in Silver Spring, Maryland. Some members of the deaf community, for example, welcome the thought of having children who are deaf, and might be concerned that ways to edit deafness mutations from the genome would increase pressure on families to do so.

Public surveys often find support for heritable genome editing — if it is shown to be safe and used to treat genetic diseases. A UK survey conducted by the Royal Society found that 83% of participants were in favour of editing the germ line to treat incurable disease. But many drew the line at editing for 'enhancement': 60%, for example, were opposed to the idea of using heritable gene editing to improve intelligence.

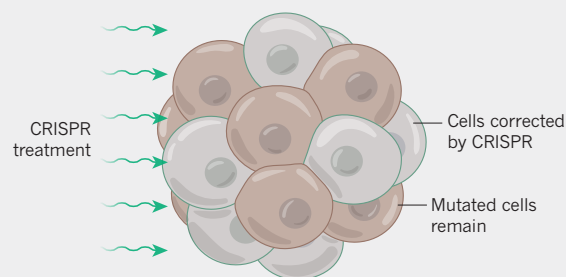
Many scientists and ethicists make a similar distinction, between modifying the genome to enhance athletic ability, for example, or to change

Mosaicism

Mosaicism can cause two sorts of problems. If a developing embryo contains just a few cells with risky mutations, then a biopsy that picks up a mutated cell might lead to unnecessary manipulations.



The CRISPR–Cas9 process is inefficient, and might leave too many cells uncorrected to treat the disease.



eye colour, versus treating or preventing disease. And even then, there is debate about which diseases might warrant such an approach. Fatal conditions with a strong, clear-cut genetic contribution — such as Huntington's disease, which is almost inevitable when the mutation is present — are the most common examples given. But when it comes to editing a gene such as *PCSK9* to prevent high cholesterol and potentially stave off heart disease, things are decidedly more grey, says Feng. Ultimately, Porteus hopes to see a registry of conditions that have been evaluated by specialists and deemed worthy of intervention with heritable gene editing, much as the United Kingdom now maintains for PGD.

Still, some people might be quietly moving towards the idea of more gene-edited children. This month, a Russian scientist announced his interest in pursuing a project to edit the genes of human embryos. And the US media company STAT reported late last month that a fertility clinic in Dubai had reached out to He for advice on gene editing shortly after he made his announcement.

Abha Saxena, a bioethicist at the University of Geneva, Switzerland, and former adviser to the WHO, hopes that consultations will continue, even if the ultimate goal of reaching a global consensus might not be possible. "Are we ever going to be ready? It's difficult to say," Saxena says. "But humanity has always been adventurous." ■

Heidi Ledford is a senior reporter for *Nature* in London, UK.

1. Kleinstiver, B. P. *et al. Nature* **529**, 490–495 (2016).
2. Ma, H. *et al. Nature* **548**, 413–419 (2017).
3. Egli, D. *et al. Nature* **560**, E5–E7 (2018).
4. Strecker, J. *et al. Science* <https://doi.org/10.1126/science.aax9181> (2019).
5. Klompe, S. E., Vo, P. L., Halpin-Healy, T. S. & Sternberg, S. H. *Nature* <https://doi.org/10.1038/s41586-019-1323-z> (2019).
6. Komor, A. C., Kim, Y. B., Packer, M. S., Zuris, J. A. & Liu, D. R. *Nature* **533**, 420–424 (2016).
7. Zuo, E. *et al. Science* **364**, 289–292 (2019).
8. Jin, S. *et al. Science* **364**, 292–295 (2019).
9. National Academies of Sciences, Engineering, and Medicine. *Human Genome Editing: Science, Ethics, and Governance* (National Academies Press, 2017).
10. Wei, X. & Nielsen, R. *Nature Med.* **25**, 909–910 (2019).
11. Lander, E. *et al. Nature* **567**, 165–168 (2019).

COMMENT

SUSTAINABILITY Manage rivers in a changing climate with process models **p.301**

GEOSCIENCE A paean to carbon-14 and the people who track its traces **p.304**

EQUITY Racism report — academics at University of Cape Town reply **p.307**



OBITUARY Murray Gell-Mann, particle predictor, remembered **p.308**

EDUARDO SOTERAS/AFP/GETTY



Students in a laboratory at the Catholic University of Graben in Butembo, Democratic Republic of the Congo.

Build science in Africa

To cope with climate change and population growth, the continent urgently needs more home-grown researchers, argue **Anagaw Atickem**, **Nils Chr. Stenseth** and colleagues.

Africa's population is projected to nearly quadruple over the next century¹. And that is following a staggering increase over just seven decades — from 200 million people in 1950 to 1.25 billion in 2018 (ref. 2). Meanwhile, temperatures across the continent are expected to rise by between 3 °C and 4 °C over the next century, bringing more drought, flooding, conflict and species loss³.

To face these formidable challenges,

Africa must improve its capabilities in higher education and research. Yet the quality of the scientific education provided at many universities on the continent has, if anything, deteriorated over the past two decades.

When two of us (A.A. and A.M.) moved from Addis Ababa University in Ethiopia to begin our PhDs in ecology and population genetics at the University of Oslo in Norway (in 2007 and 2012), we had never

set foot in a genetics laboratory before, nor even seen a PCR machine. We had taken courses in statistics while pursuing master's degrees in Ethiopia, but had never touched a computer as part of our training. Thus, we had no practical experience in complex data analysis, and no idea about how to use software programs — such as the statistical packages R or ArcGIS — that are now common tools even in undergraduate courses in ecology, genetics and conservation ►

► biology in high-income countries.

Between us, we have experience in conducting research or mentoring researchers across sub-Saharan Africa (Botswana, Ethiopia, Kenya, Rwanda, Senegal, South Africa, Tanzania and Uganda) and in other parts of the world, including China, Canada, Germany, Norway, the United Kingdom and the United States. (Four of us were born in Africa, and trained there as well as in Europe and Canada; the rest of us are from high-income countries.) Drawing on this experience, we lay out the challenges of getting training in the biological and environmental sciences as an African student — at home and abroad.

In our view, the improvements so badly needed must happen mainly in Africa's higher-education system of public and private universities. But support from the international scientific community is essential. Global research and global stability stand to benefit.

BOOM OR BUST?

During the last quarter of the twentieth century, national governments and the World Bank prioritized primary education over higher education. From 1980 to 1996, for instance, of the total budget for education in sub-Saharan Africa, 49% went to primary education and only 18% to higher education⁴.

In the early 2000s, governments and international agencies changed course, recognizing that the global economy of the twenty-first century would be driven by knowledge. A flurry of expansion in higher education across Africa followed⁵. In Ethiopia, for instance, between 2003 and 2012, the share of government expenditure on education was stable for primary education (at 35%); for higher education, it increased from 28% to 43% (ref. 6).

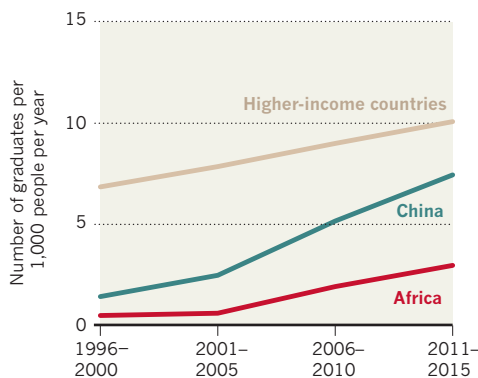
The number of public and private universities has soared. Ethiopia, for example, now has 46; before 2000, it had 2. And between 2000 and 2013, total enrolment in higher education on the continent doubled to 12.2 million⁷.

Yet most of Africa's universities lack well-equipped laboratories, libraries and other basic infrastructure, such as a reliable electricity supply or Internet connection. There is little funding for research (see 'Scant spending'). At Addis Ababa University, which ranks 37th among the best African universities in the 2018 uniRank scale, each PhD student receives about US\$16,000 in total for 4 years of study. Even accounting for the differences in the cost of living, that's about 60% less than a typical PhD student receives in Europe or the United States. Faculty members are typically poorly paid and insufficiently trained. And many focus almost exclusively on lecturing because it is almost impossible to provide students

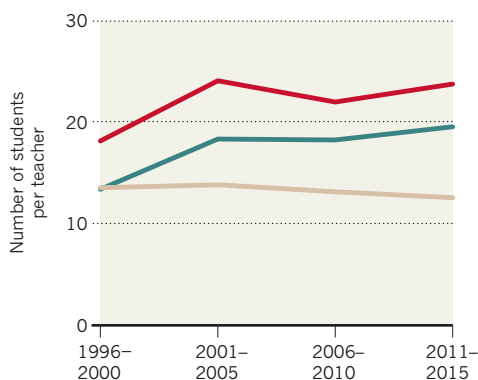
TRENDS IN HIGHER EDUCATION

Graduation from African universities is soaring, but more students per teacher and decreased spending per student mean the quality of education is likely to fall.

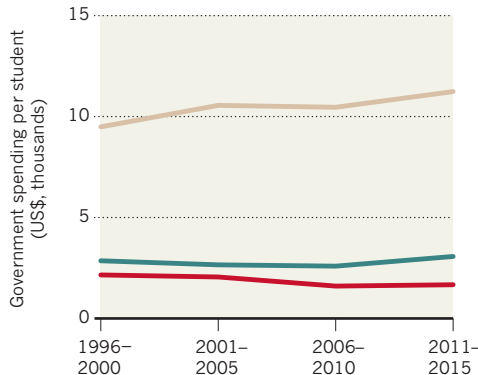
MORE GRADUATES



LESS SUPERVISOR CONTACT



SPENDING PER STUDENT DECREASING



with hands-on research experience. What's more, students have little access to scientific publications other than those that are freely available online.

To make matters worse, most of the faculty members who advise graduate students have little teaching, research or advisory experience themselves. In Ethiopia in 2010, less than 20% of university instructors held master's degrees, and less than 4% held PhDs⁸. Even in South Africa, the country with the highest total spending on research and development, only 39% of academic staff hold PhDs⁹.

This lack of training becomes a vicious circle. Graduate students from many

universities produce theses that fall short of international standards, and rarely publish their work in high-impact journals.

As university enrolment in Africa continues to mushroom, its historically underfunded higher-education systems are being stretched to breaking point⁵. Across the continent, academics now teach more classes containing greater numbers of students than ever, leaving little time for research (see 'Trends in higher education'). Poorly paid faculty members often have second jobs (at other universities, in the private sector or with non-governmental agencies), which compounds the problems. In Nigeria, for example, it is not uncommon for an academic at one university to lecture at several others. Some courses are crammed into a couple of 12-hour days to enable lecturers to get to their next teaching post.

Despite these formidable obstacles, it is an exciting time for higher education and scientific research in Africa. Decades of investment in primary and secondary education have created more candidates for further education. And the growing number of people with Internet access is making collaborations much easier, both in Africa and with the rest of the world.

OVERSEAS SUPPORT

Currently, overseas support for African scientists comes mainly from national and multinational agencies such as the World Bank, from European countries such as France, Germany, Norway, Sweden and the United Kingdom, and from individual institutions and research mentors.

Since 2014, the World Bank has provided \$500 million to build or strengthen 46 African Higher Education Centers of Excellence for postgraduate education and applied research (see *Nature* **561**, 16; 2018). So far, nearly half of these have attained international accreditation¹⁰ (see also go.nature.com/2kgcg9x). In West and Central African countries alone, 8,000 master's and 2,000 PhD students have received training at these centres so far¹⁰.

This and other 'donor-led' initiatives by the European Union and the United Nations are too fleeting. Funding often lasts for only five years or so, but needs to be provided over decades. Current trends to shrink long-term foreign aid are therefore worrying: in March, US President Donald Trump proposed a 23% cut in diplomatic and aid programmes of the type that would be needed to support intellectual growth in Africa¹¹.

When it comes to nations supporting science in Africa, there is considerable variability in approach. Among high-income host countries, France had the most students from Africa in 2016 — around 95,000 (nearly 50% in master's courses and 8%



People sell food in Mbandaka, Democratic Republic of the Congo. Research into food security will help Africa face the challenge of feeding its growing populations.

in PhD programmes)¹². And China is the fastest-growing destination for African students; according to Chinese ministry figures, it hosted around 62,000 in 2016, up from fewer than 2,000 in 2003 (see also Supplementary Information).

Since 1961, Sweden has provided long-term funding and mentoring to research groups at universities in low-income countries — mainly in Africa — in physics, chemistry and mathematics. Likewise, Norway has offered various scientific training programmes since 1970. Most notable is the Norwegian Partnership Programme for Global Academic Cooperation (NORPART), which from late 2017 replaced the Norwegian Quota Scholarship Scheme.

In 2018, NORPART awarded \$650,000 to three of us (N.C.S, A.B. and A.A.) to establish a mutually beneficial educational exchange programme between the University of Oslo and Addis Ababa University. This will promote the exchange of ideas between faculty members at each institution, and provide training to 26 Ethiopian and 15 Norwegian MSc students over the next 5 years.

Indeed, a crucial lesson from the Scandinavian experience is the importance of sustained investment over many decades for promoting self-sustaining, high-quality education and research in Africa.

Institutions in wealthier countries can also have important roles in growing scientific capacity in Africa. But too few regularly commit financial resources to this, or encourage their faculty members to get involved. And between-country connections generally rely on existing ties between individuals. These can be fragile and subject to the conditions of the granting agencies. University staff in Africa who have no contacts at a university in Norway cannot apply for NORPART funding, for example.

To be fair, mentoring students who are far from their home countries is demanding. As well as often lacking experience in

lab work, techniques in data analysis, writing research papers and so on, the students face cultural and language barriers. Most assistance programmes cover only an individual's costs, not those of dependants. But many African graduate students are older than most of their classmates from high-income countries and have families to support. Students can opt to leave their families at home, but this brings its own challenges.

THREE STEPS TO CHANGE

Incentivize faculty members in high-income countries. In our view, establishing more one-on-one mentorships for students with advisers from wealthier countries would most rapidly increase the rate at which science develops in Africa¹³.

Existing reward structures in academia favour individuals who win big grants and publish in the most prestigious journals. We think that such reward structures should adapt to meet the changing realities of a global society. Scientists who invest in international collaborations with African scientists and others from low-income countries could be given additional grants specifically for intensive mentoring, for instance. (Imagine how the rate of scientific development in Africa would change if every paper published by an African graduate student mentored by a researcher from

SCANT SPENDING

Africa devotes a tiny proportion of its GDP to research (data between 2009 and 2013).

North America and Western Europe

2.4% GDP

East Asia and the Pacific

1.9%

Sub-Saharan Africa

0.4%

Regional categories are those used by the United Nations Educational, Scientific and Cultural Organization. GDP, gross domestic product.

a high-income country counted twice as much for tenure and promotion as a paper by a graduate student from the adviser's own university.)

Support MSc and PhD students. Institutions in high-income countries that support graduate students from Africa should offer modified programmes designed to meet their needs. These could include peer-mentoring programmes, and budgeting for students to do an extra year or more courses to enable degree completion.

Support scientists in Africa. Universities across Africa need to attract and retain more qualified educators by improving working conditions. This would support all academic researchers, but especially those early in their careers. It is relatively easy to obtain a faculty job in many African countries because of the increase in the number of universities¹⁴. But pay for academic scientists is typically abysmal¹⁵. (For instance, the gross monthly salary for a new PhD holder at Addis Ababa University is less than \$500, around the same as the monthly rent for a typical two-bedroom apartment nearby.) Also, the limited facilities and support for research in sub-Saharan Africa is dismaying for newly minted African PhDs¹³.

Many African PhD holders returning home after a stint abroad find it difficult — or impossible — to remain competitive in their field¹³, or to support a family on the salary provided. Instead, they opt for more lucrative work in the private sector or in international development. Others never return, contributing to Africa's 'brain drain'^{16,17}.

To encourage African PhDs to stay in

2.6%

of the world's peer-reviewed scientific journal articles in 2013 included at least one author based in Africa.

science in Africa, foundations and other granting bodies worldwide should offer more multi-year research grants that support both basic and applied research on the continent. Funding schemes, similar to those in China and some European countries, should be developed to incentivize African students trained in high-income countries to contribute to scientific development in their country of origin (see *Nature* **569**, 325–326; 2019).

WHO BENEFITS?

Developing science in Africa will improve global political and economic stability. And research as a whole will be strengthened — by different ideas born of different challenges and experiences — if many more African scientists become key members of the global scientific community.

Ultimately, it must be Africans themselves who drive the transition towards a stronger Africa. Wealthy African entrepreneurs could help with this. Governments across the continent must bring their own money to higher education and research. But improving Africa's higher-education and research capacity will also require considerable spending by — and structural change in — the academic and financial institutions of richer nations. Most

importantly, scientists working in high-income countries or emerging economies need to change the way they view, value and reward collaborations involving researchers from Africa. It is only by working together that sustainable top-level African universities will develop — institutions that are needed if we are to tackle the major global issues of our time. ■

Anagaw Atickem is an adjunct assistant professor in the Department of Zoological Sciences at Addis Ababa University, Ethiopia, and is associated with the Centre for Ecological and Evolutionary Synthesis, University of Oslo, Norway. **Nils Chr.**

Stenseth is professor at the Faculty of Mathematics and Natural Sciences and at the Centre for Ecological and Evolutionary Synthesis at the University of Oslo, Norway; honorary professor in the Department of Zoological Sciences, Addis Ababa University, Ethiopia; and an elected member of the World Academy of Sciences. **Peter J.**

Fashing, Nga Nguyen, Colin A. Chapman, Afework Bekele, Addisu Mekonnen, Patrick A. Omeja, Urs Kalbitzer.

e-mail: n.c.stenseth@ibv.uio.no

1. United Nations Department of Economic and Social Affairs. *World Population Prospects: The 2015 Revision. Key Findings and Advance Tables* (UN, 2015).
2. Altenburg, T. et al. *Foresight Africa: Top Priorities for the Continent in 2019* (Brookings Inst., 2019).
3. Malhi, Y., Adu-Bredu, S., Asare, R. A., Lewis, S. L. & Mayaux, P. *Phil. Trans. R. Soc. B* **368**, 20120312 (2013).
4. Stasavage, D. *Am. J. Polit. Sci.* **49**, 343–358 (2005).
5. World Bank. *Accelerating Catch-up: Tertiary Education for Growth in Sub-Saharan Africa* (World Bank, 2009).
6. World Bank. *Ethiopia Public Expenditure Review 2015* (World Bank, 2016).
7. UN Educational, Scientific and Cultural Organization. *UNESCO Institute for Statistics: Higher Education* (2018).
8. Reisberg, L. & Rumbley, L. E. *Int. Higher Edu.* **45**, 23–24 (2010).
9. Cloete, N., Sheppard, C. & Bailey, T. in *Knowledge Production and Contradictory Functions in African Higher Education* (eds Cloete, N., Maasen, P. & Bailey, T.) 75–108 (African Minds, 2015).
10. World Bank. *African Centers Of Excellence Project (P126974)* (World Bank, 2018).
11. Rabinowitz, K. & Uhrmacher, K. 'What Trump proposed in his 2020 budget' *The Washington Post* (11 March 2019).
12. Marshall, J. 'International African students in France — A Profile' *University World News* (17 November 2016).
13. Dike, V. N. et al. *Nature Clim. Change* **8**, 447–449 (2018).
14. Teferra, D. J. *High. Edu. Africa* **11**, 19–51 (2013).
15. UN Educational, Scientific and Cultural Organization. *UNESCO Science Report 2010: The Current Status of Science Around the World* (UNESCO, 2010).
16. Watkins, A. & Mandell, J. *Global Forum Action Plan: Science, Technology and Innovation Capacity Building Partnerships for Sustainable Development* **32** (World Bank, 2010).
17. Capuano, S. & Marfouk, A. J. *Comp. Policy Anal. Res. Practice* **15**, 297–314 (2013).

Supplementary information accompanies this article: see go.nature.com/31sxhsx.



The class of 2015 graduating from the American University of Nigeria in Yola.



Dead fish on the banks of the Guadiaro River in southern Spain during severe drought.

Prepare river ecosystems for an uncertain future

As the climate warms, we can't restore waterways to pristine condition, but models can predict potential changes, argue **Jonathan D. Tonkin**, **N. LeRoy Poff** and colleagues.

In January, millions of fish died in Australia's Murray–Darling Basin as the region experienced some of its driest and hottest weather on record. The heat also caused severe water shortages for people living there. Such harsh conditions will become more common as the world warms. Iconic and valuable species such as the Murray cod (*Maccullochella peelii peelii*) — Australia's largest freshwater fish — could vanish, threatening biodiversity and livelihoods.

Rivers around the world are struggling to cope with changing weather patterns. In Germany and Switzerland, a heatwave last year killed thousands of fish and blocked shipping on the River Rhine. California is emerging from a six-year drought¹ that restricted water supplies and devastated trees, fish and other aquatic life. Across

the US southwest, extended dry spells are destroying many more forests and wetlands.

What should river managers do? They cannot look to tools of old: conventional management techniques that aim to restore ecosystems to their original state. Ongoing human development and climate change mean that this is no longer possible. And models based on past correlations do a poor job of predicting how species might respond to unprecedented changes in future (see 'Ecosystem change'). A different approach is called for.

To maintain water supplies and avoid devastating population crashes, rivers must be managed adaptively, enhancing their resilience and limiting risk. Researchers must also develop better forecasting tools that can project how key species, life stages and

ecosystems might respond to environmental changes. This will mean moving beyond simply monitoring the state of ecosystems to modelling the biological mechanisms that underpin their survival.

MODEL PROCESS

Today, river managers track properties such as species diversity and population abundance, and compare them with historical averages. If they spot troubling declines, they might intervene by, for instance, altering the amount of water released from dams. But by the time trends are detected, they can be impossible to arrest.

Understanding how sensitive ecosystems might change is crucial to managing them in the future. For example, in the American west, native cottonwoods (*Populus* spp.) are

valuable, long-lived trees that anchor river banks and offer habitats for many species. They are finely tuned to seasonal flood patterns, releasing their seeds in early summer when river flows peak. The seeds take root in moist ground after the flood recedes². But if the flood is delayed, even by a few days, many seeds fall on dry ground and die. Drought-tolerant species, such as salt cedar (*Tamarix ramosissima*), that disperse seeds over a longer period will move in and dramatically alter conditions for native flora and fauna.

Models based on biological processes or mechanisms — that is, how rates of survival, reproduction and dispersal vary with environmental conditions — can follow and predict such shifts. For example, by modelling the impacts of changes in flood timing on aquatic invertebrates, it is possible to predict how the numbers of dragonflies and mayflies in a dryland river will vary with different patterns of dam releases³.

Process-based models can be tailored to particular life stages of a species, or sequences of events⁴. They can identify tipping points and bottlenecks. For example, they have revealed that the early juvenile stage of coho salmon (*Oncorhynchus kisutch*) in the northwestern United States is most sensitive to summer droughts. The salmon spawn in streams that flow into coastal rivers, and might spend a couple of years in fresh water before moving to the sea. Juveniles might not survive, or might find it hard to travel downstream, when the river levels are low⁵.

Such models can also track how interactions among species in communities vary under changing conditions⁶. For example, the loss of riparian specialists in dryland river ecosystems and invasion by both non-native and upland species in a drier future could create a vicious cycle⁶. River ecosystems could become more vulnerable to climate change and to alien species.

Armed with all this information, managers can intervene before a problem arises. For example, in wet years, conservationists in the Pacific Northwest could find and support habitats that are crucial to juvenile salmon.

They could manage water flows in dry years to enable the salmon to migrate. Similarly, in the US southwest, river flows could be increased strategically from reservoirs to protect important species, such as cottonwoods. And in Australia, letting more water pass through dams in spring could stop rivers drying up while the eggs of Murray cod mature⁷.

Rivers must also be managed for people. Allocating scarce water resources is contentious. Policymakers, water-resource engineers, conservationists and ecologists must work together to decide how much water should be diverted to people, agriculture and industry, and how much is needed to protect ecosystems during drought.

“Dam managers should focus on the most vulnerable or responsive life stages.”

Some river basins are beginning to be managed adaptively — agencies are trying different management practices, learning from them and updating them as needed. For example, in Australia, state and federal agencies periodically reassess and rebalance water allocations, as climate trends, information and assessment tools develop. Similarly, the Bay-Delta Plan in California proposes to revisit relationships between target species, water flows and water quality in San Francisco Bay and the Sacramento-San Joaquin River Delta every five years.

But adaptive management alone might miss conservation targets. Unexpected consequences could emerge over the long term as impacts mount. Process-based models can look further ahead and save time, money and disruption by limiting the number of interventions as well as avoiding adverse impacts. They would help stakeholders and managers to choose which features of ecosystems to maintain, to justify costly interventions such as major engineering works and to weigh trade-offs to build resilience under increasing climatic uncertainty⁸.

OBSTACLES TO IMPLEMENTATION

Process-based models are already used in fisheries and conservation. For example, they have shown conservationists that it is more effective to protect juvenile loggerhead sea turtles from being caught in fishing nets than to safeguard their eggs on beaches⁹. And such models help to guide the management of wetland habitats in the

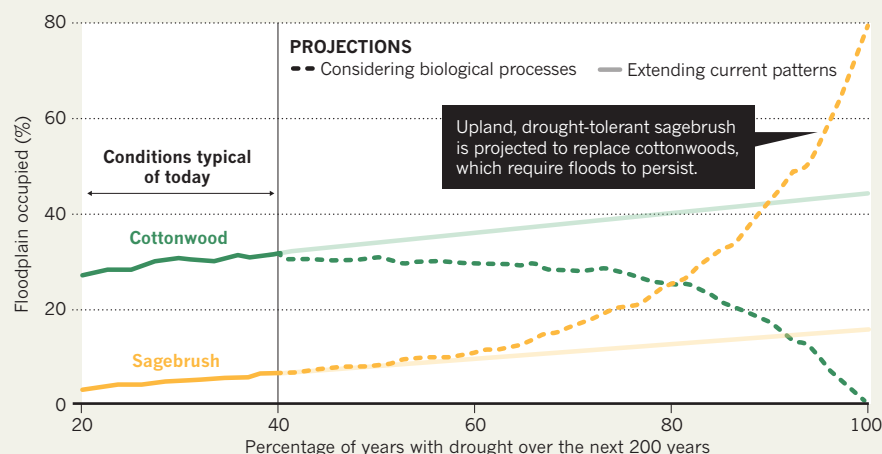


Native cottonwoods are being displaced by non-native salt cedar in the Dolores River, Utah, owing to flow alteration by damming.

MARK ULIASZ/ALAMY

ECOSYSTEM CHANGE

Models of river ecosystems must describe how various species react as droughts become more common in a warmer climate over the next 200 years. Models that consider basic biology predict broader changes than linear extrapolations miss.



United States for the endangered Everglades snail kite (*Rostrhamus sociabilis*), the fledglings of which are susceptible to droughts¹⁰.

But they are rarely used in river management, mainly because data on the basic biology of local species are lacking. Such data are costly for scientists and agencies to collect. Measuring fecundity or survival, for example, takes years and thus requires long-term funding and commitment. Such campaigns are usually reserved for endangered or commercially valuable species.

Simplifying models might help to bridge the data gaps in the interim. Species with similar life histories or characteristics might respond similarly to changing river conditions. Studies of one could inform models and management of similar species in other places. For instance, plains cottonwood (*Populus deltoides*) in North America, river red gum (*Eucalyptus camaldulensis*) in Australia, and Euphrates poplar (*Populus euphratica*) in North Africa and Eurasia are all riparian trees that have similar hydrological requirements and drought tolerances. They share characteristics such as shallow roots and furrowed bark that resists flood scour, and can resprout after being buried by sediment. Analytical methods could also be developed to extrapolate across gaps in data sets.

FOUR STEPS

River scientists and managers should take the following steps.

Collect data on mechanisms. We call for a fresh global campaign to gather natural-history data on the responses of biodiversity to changes in river flow. Estimates of fecundity and survival at various life stages will require monitoring in the field. Other information, such as flood-induced mortality rates, could be gathered through field and

laboratory experiments. Data from different sources can also be combined, including species traits, population abundances across life stages and remote-sensing data about the states of ecosystems on wider scales⁴.

We urge local, state and federal agencies, as well as researchers, non-governmental organizations and other bodies, to make existing data available. Facilities for hosting these already exist, such as the COMPADRE and COMADRE global databases, which hold population models for hundreds of plant and animal species, respectively. Organizations such as the Alliance for Freshwater Life, the wildlife charity WWF and the Group on Earth Observations Biodiversity Observation Network should lobby global funding bodies to support data collection.

Describe key processes in models. Scientists need to better articulate the relationships between population dynamics and water-flow patterns in process-based models. For example, the models need to describe how well different life stages of plants reproduce or survive under flood or drought conditions, the flow conditions and timing that are required for fish to reproduce or the growth rates of insect populations after floods of different sizes^{3–5,7}. Outputs need to be expressed clearly so that river managers and decision makers can understand and use them.

Focus management on bottlenecks. Targeted interventions to avoid populations collapsing during extreme flows will be a cornerstone of managing rivers for resilience in future. Accordingly, dam managers should focus on the most vulnerable or responsive life stages, not just population abundance. Sadly, as flow extremes become more common, scientists and managers will be able to observe die-offs and calibrate the models.

Pinpoint uncertainty. The level of confidence that managers have in the results of models will influence how willing they will be to deal with varying levels of risk. Predictions should thus quantify the level of trust that can be placed in them. Scientists must present uncertainties in forecasts clearly. Models should be tested by hind-casting (predicting past or present population size, for example), and uncertainty in model inputs should be traced through to the outputs. The knowledge gaps that most compromise accuracy should be identified. The models should be regularly updated, tested and improved as new data arrive.

Freshwater biodiversity is disappearing on our watch. As the crisis deepens, we must model and manage rivers to safeguard the services they provide. ■

Jonathan D. Tonkin is a Rutherford Discovery Fellow in the School of Biological Sciences at the University of Canterbury, Christchurch, New Zealand. **N. LeRoy Poff** is a professor in the Department of Biology at Colorado State University, Fort Collins, USA, and at the Institute for Applied Ecology, University of Canberra, Australia. **Nick R. Bond**, **Avril Horne**, **David M. Merritt**, **Lindsay V. Reynolds**, **Julian D. Olden**, **Albert Ruhi**, **David A. Lytle**.

e-mails: jonathan.tonkin@canterbury.ac.nz; n.poff@colostate.edu

1. Diffenbaugh, N. S., Swain, D. L. & Touma, D. *Proc. Natl Acad. Sci. USA* **112**, 3931–3936 (2015).
2. Mahoney, J. M. & Rood, S. B. *Wetlands* **18**, 634–645 (1998).
3. McMullen, L. E., De Leenheer, P., Tonkin, J. D. & Lytle, D. A. *Ecol. Lett.* **20**, 1566–1575 (2017).
4. Lytle, D. A., Merritt, D. M., Tonkin, J. D., Olden, J. D. & Reynolds, L. V. *Ecol. Appl.* **27**, 1338–1350 (2017).
5. Ohlberger, J. et al. *Freshwater Biol.* **63**, 331–340 (2018).
6. Tonkin, J. D., Merritt, D. M., Olden, J. D., Reynolds, L. V. & Lytle, D. A. *Nature Ecol. Evol.* **2**, 86–93 (2018).
7. Yen, J. D. L., Bond, N. R., Shenton, W., Spring, D. A. & Mac Nally, R. J. *Appl. Ecol.* **50**, 691–701 (2013).
8. Poff, N. L. et al. *Nature Clim. Change* **6**, 25–34 (2016).
9. Crouse, D. T., Crowder, L. B. & Caswell, H. *Ecology* **68**, 1412–1423 (1987).
10. Beissinger, S. R. *Ecol. Appl.* **5**, 618–631 (1995).

CORRECTIONS

The Comment ‘Make scientific data FAIR’ (*Nature* **570**, 27–29; 2019) wrongly stated that Springer Nature has signed up to the Enabling FAIR Data Project’s Commitment Statement; so far, only *Nature* and *Scientific Data* have done so. The Comment ‘Credit data generators for data reuse’ (*Nature* **570**, 30–32; 2019) wrongly located Julie Dunning Hotopp at the University of Maryland in College Park; in fact, she is in Baltimore.



A human femur, thought to be from medieval times, being sampled for carbon dating.

GEOSCIENCE

Radiocarbon revolution

Chris Turney applauds a book on carbon-14 and its key applications.

It is nearly 80 years since the discovery of carbon-14, a radioactive isotope of the sixth element. Because its decay can be used to track the passage of time, radiocarbon has made myriad contributions across the Earth, environmental, biological and archaeological sciences. In the wonderfully engaging *Hot Carbon*, oceanographer John Marra takes this story much further, exploring not just the science, but why we should care about it.

Radiocarbon is scarce in nature, formed in the upper atmosphere through the interaction of cosmic rays with nitrogen. It is rapidly converted to carbon dioxide, and filters into a host of carbon reservoirs in the biosphere and ocean. Living organisms constantly take up ^{14}C , and after they die, the isotope decays at a known rate. By measuring the amount left in a carbon-based sample, it is possible to calculate its age. Since the 1940s, the technique has been used to date materials as much as 60,000 years old, capturing everything from the early migration of modern humans out of Africa, by dating bones and charcoal from ancient hearths, through to the incredibly slow growth rates of mosses living on the fringes of Antarctica. In retelling these facts, Marra offers compelling stories about the great researchers — many long forgotten

— whose discoveries made possible the theory, practice and further findings we now take for granted. There's enough to satisfy the most insatiable informavore.

Hot Carbon starts with the extraordinary story of chemist Martin Kamen, born in Canada to Russian immigrants. In February 1940, Kamen was trying to produce a new isotope of carbon at the Berkeley Radiation Laboratory at the University of California. Sleep-deprived after three nights of collecting sufficient irradiated graphite to measure the hoped-for isotope, he stepped outside. His bedraggled appearance caught the attention of police; worse, he fitted the description of an escaped convict who had gone on a murder spree. Hauled to the police station, Kamen was finally released when a survivor of the bloodbath confirmed he was not the suspect. Kamen returned to the laboratory to find that his colleague Sam Ruben had analysed



**Hot Carbon:
Carbon-14 and
a Revolution in
Science**

JOHN F. MARRA
Columbia University
Press (2019)

the carefully gathered sample and found that it was measurably radioactive. The story of ^{14}C thus began with a dose of high drama.

Originally expected to have a half-life of just minutes or hours, this heavy form of carbon was considered a low research priority. But Kamen and Ruben's efforts proved that it would be stable over millennia, opening up a breathtaking number of research avenues (its half-life of 5,730 years was determined some years later). Kamen never received the credit he deserved, becoming a victim of the US anti-communist fervour of the 1940s and 1950s. Those who applied his insight, such as chemists Willard Libby and Melvin Calvin, reaped the scientific reward.

We follow the ^{14}C trail through a number of disciplines, learning, for instance, how Calvin and his team used the isotope to trace the way in which plants convert CO_2 into sugar, revealing the intricate processes underpinning photosynthesis. We see how radiocarbon was deployed by labs in Britain, Switzerland and the United States to date the flax used to weave the Turin Shroud (believed by some to be the burial cloth of Jesus) to between 1260 and 1390. Radiocarbon dating has shown that Ötzi — the corpse retrieved from melting alpine ice on the Austrian-Italian border in 1991 — is more than 5,000 years

old. And we discover how candidate drugs, labelled with ^{14}C at specific parts of the molecule, can be followed through phases of the body's metabolism to test the drugs' safety and efficacy. There is so much more. Marra explains, for instance, how, shortly after ^{14}C was discovered, dissolved CO_2 in seawater was used to track the movement of currents in the deep ocean, revealing connections around the planet considered unfathomable before.

Carbon-14 may be the star, but scientists, institutions and happenstance have valuable supporting roles. Take Libby, winner of the 1960 Nobel Prize in Chemistry for his work developing radiocarbon dating. At one point, his team waded into the sewers of Baltimore, Maryland, collecting methane produced from human waste to demonstrate unequivocally that it contained considerably more ^{14}C than did archaeological samples and a precisely dated piece of redwood heartwood.

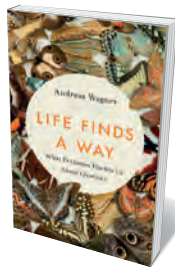
Marra also reveals, in vivid detail, the difficulties faced by early researchers in acquiring precious samples of plankton, which opened up a new perspective on ocean productivity and, ultimately, carbon sequestration. His own experience in this area illuminates the researchers' pioneering spirit in the face of wild conditions, cramped spaces and sometimes surly ships' captains. The technological limitations were progressively overcome by dogged perseverance and a belief that the work would help them to understand the oceans' potential for incorporating inorganic carbon into organic compounds — still the focus of fierce investigation.

Mysteries remain in the Earth sciences, such as the effectiveness of the carbon cycle and the ramifications of human activity, including our seemingly insatiable hunger for fossil fuels. Importantly, Marra shows how ^{14}C can be used to tease out processes across a range of timescales. He explains why the Southern Ocean is the 'gatekeeper' to the planet's ocean circulation, and how abrupt changes in the formation of deep water and the position of the overlying wind belts can drive dramatic shifts in the carbon cycle. Soberingly, a doubling of atmospheric levels of ^{14}C — arising from mid-twentieth-century nuclear-bomb testing — is preserved as a spike in annually formed natural archives, including tree rings. That marker could be chosen to delineate the start of a new geological epoch: the Anthropocene.

Hot Carbon offers a timely perspective on how mind-bogglingly connected our planet is — and how ^{14}C will continue to be important in helping us to understand what lies ahead. ■

Chris Turney is professor of climate change and Earth science, and director of the Chronos ^{14}C -Carbon-Cycle Facility, at the University of New South Wales in Sydney, Australia. His website is christurney.com e-mail: c.turney@unsw.edu.au

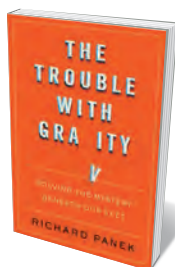
Books in brief



Life Finds a Way

Andreas Wagner BASIC (2019)

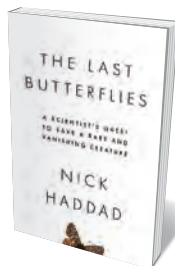
From gut microbes to elephants, the embodiments of evolution are marvels of innovation. So writes biologist Andreas Wagner, whose eloquent study finds the “augmented” view of Darwinian evolution echoed in the ‘landscape thinking’ of human creativity — the mental exploration of possibility. Wagner meshes research into areas such as genetic drift with theories on aspects of the creative process (such as serendipity) seen in luminaries from radioimmunoassay inventor Rosalyn Yalow to artist Pablo Picasso, and shows how such a mindset can solve real-world problems.



The Trouble with Gravity

Richard Panek HOUGHTON MIFFLIN HARCOURT (2019)

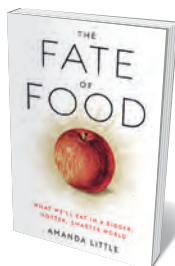
Gravity is, scientifically speaking, an enigma. We know that it is a weak force with infinite reach, pulling at matter to form bodies from black holes to galaxies — but that is what it does, not what it is. Science writer Richard Panek follows the evolution of gravitational theories through time, from Aristotle (arguably the first to question why things fall on Earth but heavenly bodies stay aloft) through the momentous discoveries of Galileo Galilei, Isaac Newton and Albert Einstein. He also touches on insights from the likes of poet Dante Alighieri and philosopher-physicist Ernst Mach.



The Last Butterflies

Nick Haddad PRINCETON UNIVERSITY PRESS (2019)

Terrestrial ecologist Nick Haddad studies the beleaguered denizens of liminal lands: the world's rarest butterflies. Far from niche research, he argues, findings on species and subspecies such as the St Francis' satyr (*Neonympha mitchellii francisci*) and crystal skipper (*Atrytonopsis quinteri*) offer a valuable lens on the biodiversity crisis. Yet Haddad does not just gather data on habitat loss and other drivers of decline — although he does that with crystalline acuity. He emphasizes that measures such as restoring ecological systems can protect populations of these fragile “ambassadors of nature”, against the odds.



The Fate of Food

Amanda Little HARMONY (2019)

As Earth's human population tips towards 8 billion, is our food system up to it? In this tour of the brave new world of adaptive production, environmental journalist Amanda Little encounters robot weeders, aquaponics innovators, permaculture farmers, toilet-to-tap water recyclers and other players in the field — both high-tech innovators and eco-traditionalists. That balanced viewpoint extends to broader discussions of debates over genetically modified crops in Kenya and cultured meat production in the United States. Witty, nimble and timely, this is a gem of crack reporting.



Giants of the Monsoon Forest

Jacob Shell W. W. NORTON (2019)

In this thought-provoking study, geographer Jacob Shell probes an unusual interspecies alliance: the relationship between people and the working elephants of Indonesia, Myanmar and India. Never selectively bred, the night-roaming pachyderms are essentially wild, yet engage in cognitively demanding tasks such as rescue work during floods. Examining everything from the muscular miracle of the beast's proboscis to the species' wartime work, Shell also charts the threats facing Asian elephants, and the dearth of local voices in relevant policymaking. [Barbara Kiser](#)

Correspondence

Mind fixers should join forces

In her review of my book *Mind Fixers*, Alison Abbott chides me for being “pessimistic” about biological psychiatry’s prospects and failing to call attention to “new insights about brain circuitry as a potential target for treatment” (*Nature* **568**, 314–315, 2019). Her criticism misses my historical point.

The research to which Abbott refers might or might not have a transformative effect on the care of the mentally ill. Either way, we have other work to do. The ‘old’ biological psychiatry launched in the 1980s — the one under which so many of us have conducted our business for the past 40 years — is running into the sands. The confusion and distrust it has engendered is entrenched in the public sphere.

We are at an inflection point. Either we can double down or we can call for a stocktaking. What do we want — and what do patients want — from a future biological psychiatry? We need to know how to engage constructively with the distrust, and how to balance the relationship between immediate patient needs and long-term scientific insights.

I am not pessimistic at all. I am energized: crises are times of opportunity. I am hoping that *Mind Fixers* will encourage more people to discuss the choices and challenges.

Anne Harrington *Harvard University, Cambridge, Massachusetts, USA.*
aharring@fas.harvard.edu

University racism: report too gloomy

I take issue with the implication in the Institutional Reconciliation and Transformation Commission report that nothing significant has been done about racism at the University of Cape Town (UCT) since the end of apartheid in 1994 (see

Nature **568**, 151–152; 2019). As a member of the executive team during 2015–16, I can attest to the tireless efforts of management, teaching staff and students, black and white, to create conditions in which all students feel welcome and can flourish.

The student profile at UCT is beginning to reflect the demographics of the country (white students now constitute a minority). For many of us black South Africans whose first undergraduate experience was at institutions for separate races, the opportunities, support and efforts at redress are a far cry from what we had known before.

However, it cannot be said that UCT, or any other previously white tertiary institution, is free of racism. The transformation of UCT’s staff is still too slow and is becoming a matter of urgency.

‘Decolonizing’ curricula is perhaps the most complex and contested of the tasks ahead.
Anwar Suleman Mall *University of Cape Town, South Africa.*
anwar.mall@uct.ac.za

University racism: academics reply

We wish to point out that there are crucial omissions from the report by the Institutional Reconciliation and Transformation Commission (IRTC) that challenge its findings of “rife” racism at the University of Cape Town. These were not picked up in your overall assessment of the situation (see *Nature* **568**, 151–152; 2019).

For example, there is evidence that bias in the university’s promotion systems has been disappearing over the past 11 years (see H. Sadiq *et al. High. Educ.* <http://doi.org/c65d>; 2018). The profiles of academic staff and of researchers (see, for example, G. D. Breetzke and D. W. Hedding *Stud. High. Educ.* <http://doi.org/gfzdbg>; 2019)

are changing at South African universities. We also understand that one of the five members of the commission disagreed with the theoretical approach and the factual conclusions in the section of the report dealing with racism.

The commission was established as part of an agreement aimed at ending campus disruption and violence (go.nature.com/2kjdabl). It was independent of the university, but not of South African politics (go.nature.com/2kksqge). We find it unfortunate that sections of the IRTC report took a contentious approach to racism that reflected the concerns of particular political groups.

The University of Cape Town remains strongly committed to transforming its racist legacy in academic appointments and promotions (see go.nature.com/2kkdjej).

B. Daya Reddy* *University of Cape Town, South Africa.*

**On behalf of 16 co-signatories (see go.nature.com/2lav7i7 for full list).*
daya.reddy@uct.ac.za

Forests: economic perks of plantations

Strategies for global forest restoration must be framed in the context of the growing worldwide demand for fuel, food and incomes (see S. Lewis *et al. Nature* **568**, 25–28; 2019). In our view, plantations are, alongside natural and regenerating forests, a legitimate and valuable component of the global restoration strategy to directly and indirectly contribute to climate-change mitigation.

The demand for forest products is projected to increase by around 50% by 2030. Plantations are highly efficient wood-production systems and can alleviate pressures on natural forests. Along with agroforests, they also support livelihoods. The

regeneration of secondary forests, by contrast, has high ‘opportunity costs’ because the land is no longer available for more economically rewarding purposes. A fair analysis of carbon must consider continued forest degradation in restoration scenarios that exclude plantations, as well as the role of plantations in replacing steel and concrete with timber.

Plantation companies can be important investors in natural-forest restoration. New plantation models favour landscape mosaics of different species, interspersed with natural forests and agricultural lands that provide raw materials and social and environmental benefits (C. L. C. Liu *et al. Global Ecol. Conserv.* **15**, e00419; 2018).

In Brazil, for example, companies must allocate 50–80% of their lands to natural-forest restoration or protection. Members of the conservation group WWF’s New Generation Plantations collectively manage 11 million hectares, more than half of which is dedicated to conserving and restoring forests and other ecosystems, and to supporting small-scale farming (L. N. Silva *et al. New Forests* **50**, 153–168; 2019).

Jaboury Ghazoul *ETH Zurich, Zurich, Switzerland.*

Miguel Bugalho *University of Lisbon, Portugal.*

Rodney Keenan *University of Melbourne, Victoria, Australia.*
jaboury.ghazoul@env.ethz.ch

**Competing interests declared; see go.nature.com/2wrtcx6.*

CONTRIBUTIONS

Correspondence may be submitted to **correspondence@nature.com** after consulting the author guidelines and section policies at **<http://go.nature.com/cmchno>**.

Murray Gell-Mann

(1929–2019)

Theoretical physicist who won a Nobel for codifying fundamental particles.

Nobel laureate Murray Gell-Mann once described himself as “a character out of Damon Runyon”. Like the novelist’s gritty characters (they inspired the musical *Guys and Dolls*), Gell-Mann had flair, and peppered his exacting speech with slangy cracks and sarcastic put-downs. He mocked the scientific establishment’s jargon by giving his inventions jokey names. He worked incredibly diligently, but often claimed to be simply tossing off ideas. He attributed his first major insight to a slip of the tongue, and wrote the equations for two other seminal breakthroughs on napkins. He died on 24 May 2019.

Gell-Mann was a theorist in elementary particle physics. When he entered the field in the late 1940s, powerful accelerators were starting to make particles beyond the familiar proton, neutron and electron. Researchers badly needed the equivalent of a periodic table to map the relationship between all of these. Gell-Mann’s most noted contribution was to create one, and to use it to show where more new particles could be found.

Born in 1929 in Manhattan, New York City, Gell-Mann went to Yale University in New Haven, Connecticut, at 15, to study physics. In 1948, he entered the Massachusetts Institute of Technology in Cambridge, determined to earn a PhD in two years. He castigated himself for taking an extra six months to write up his thesis. In 1951, he moved to the Institute for Advanced Study at Princeton, New Jersey, to work with Robert Oppenheimer.

The next year, as a 22-year-old postdoc with Enrico Fermi at the University of Chicago, Illinois, he tackled a knotty problem. A class of recently discovered particles — produced in abundance in cosmic-ray collisions — had unexpectedly long lifetimes. “Easy come, easy go” was a tenet of the particle world. These were therefore labelled “strange” particles. In a talk at the Institute for Advanced Study, the famous slip of the tongue gave him an idea that the particles had a previously unknown fundamental property — he labelled it “strangeness”. (Others independently had a similar idea.) Strangeness was a new quantum number: something that expresses the values that certain kinds of particle can have.

At first, he was afraid to commit himself to print. A conscription notice at the end of the Korean War — his exemption paperwork had not been filed properly — prompted him to send an article to the *Physical Review*. It referred to the new “Curious Particles”, and the editors objected. The published



version concerns “New Unstable Particles” — a phrase he deemed “sufficiently pompous”.

In 1955, Gell-Mann moved to the California Institute of Technology (Caltech) in Pasadena. That year, he also married the archaeologist Margaret Dow. A few years later, he devised a scheme to codify particles, grouping all those known into eight families — the Eightfold Way, he named it in joking homage to Buddhism. Given all the other theoretical ideas flying around, few physicists paid heed. One family in Gell-Mann’s scheme had a glaring hole. At a conference in July 1962 at CERN, Europe’s particle-physics laboratory near Geneva, Switzerland, he urged experimenters to find the missing particle, naming it the Omega Minus.

At lunch with two of the experimenters, attending from Brookhaven National Laboratory in New York, he sketched out on a napkin how the particle might be found by indicating the particles into which it would decay. The two — Nicholas Samios and Jack Leitner — took the napkin back to Brookhaven, and used it to convince their director to give them high priority for running time on the lab’s accelerator. They then found the Omega Minus. It was a triumphant discovery, and it vindicated the soundness of Gell-Mann’s entire scheme. Gell-Mann called Samios and said, with his usual nonchalance: “Nick, I hear you have found something very interesting.”

The second napkin episode took place the next year, in the faculty dining room at Columbia University in New York City. Over lunch, his host Robert Serber asked Gell-Mann if the particles of the Eightfold Way were formed by mixing and matching subunits. “So I showed him why I hadn’t considered it,” Gell-Mann said. Scribbling

equations on what was to hand, he explained that such subunits would have to have fractional charges. By the time of his talk the next day, however, he’d thought, “What the hell, why not?” and proposed the idea. Reacting against “pretentious scientific language”, Gell-Mann called the subunits quarks, after a passage in James Joyce’s *Finnegans Wake*.

Meanwhile, Richard Feynman, whose office at Caltech was next to Gell-Mann’s, proposed a similar idea, calling the subunits partons. This led to a long feud in which the two vied over the name, Gell-Mann mocking Feynman’s idea as “put-ons”. When members of the physics community tried to conciliate by calling the subunits “quark-partons”, Gell-Mann prevailed.

In 1969, Gell-Mann was awarded the Nobel Prize in Physics “for his contributions and discoveries concerning the classification of elementary particles and their interactions”. For the next few decades, he continued to be a leader in developing the theory of particle physics. In those years, Samios recalled, “when I’d ask a particle theorist why they were working on something like current algebra or group theory, the answer was invariably, ‘Because Murray is working on it!’”

Given his writing habits and hypercritical temperament, few expected him to undertake a popular book. In 1994, he published one. *The Quark and the Jaguar*, he called it; the quark represented the simple side of nature, the jaguar the complex.

In talks, Gell-Mann liked to dwell, not on the triumphs of himself and others, but on the confusions, mistakes and vacillations that blocked their way. The practice was closet self-congratulation, some carped. But there was more to it. Gell-Mann would sometimes recite to audiences a ditty that he had seen on the wall of a doughnut shop:

*As you ramble on through life, Brother,
Whatever be your goal,
Keep your eye upon the doughnut,
And not upon the hole.*

Gell-Mann would add: “I try to keep my eye on the hole.” ■

Robert P. Crease is chairman of the Department of Philosophy at Stony Brook University, New York. His most recent book is *The Workshop and the World: What Ten Thinkers Can Teach Us About Science and Authority* (Norton).
e-mail: robert.crease@stonybrook.edu

Sunlight harvested by nanotubes

Rolling an atom-thick semiconductor layer into a nanoscale tube allows it to convert solar energy into electricity without the need for semiconductor junctions — prerequisite features of conventional solar cells. [SEE LETTER P.349](#)

MING-MIN YANG & MARIN ALEXE

For decades, the development of a cheap and efficient way to convert sunlight into electricity has been at the forefront of research, from the physical sciences to engineering. Usually, devices for harvesting solar energy, called solar cells, are made of semiconductors such as silicon. In these devices, electrical power is generated at the junction between two types of semiconductor material. However, the efficiency of junction-based solar cells has almost reached its theoretical limit, and it is therefore imperative to explore methods for converting sunlight into electricity that do not require semiconductor junctions. On page 349, Zhang *et al.*¹ report a key advance in this direction. They demonstrate a junction-free solar cell that is produced by curling an atom-thick semiconductor layer into a nanoscale tube.

In a conventional solar cell, different chemical elements are added to two regions of a semiconductor in a process known as doping. Electrical transport occurs through negatively charged electrons in one region and through positively charged electron vacancies, called holes, in the other. An electric field is generated at the junction between these two regions. When sunlight is absorbed at this junction, electron-hole pairs are produced. The electrons and holes are then separated by the electric field, giving rise to an electric current (Fig. 1a). This conversion of solar energy into electricity is known as the photovoltaic effect.

Zhang and colleagues fabricated junction-free solar cells using the semiconductor tungsten disulfide. Crystals of this material have a layered structure, and can be peeled off layer by layer in a similar way to graphite. The resulting atom-thick sheets can then be rolled by chemical methods into tubes that have diameters of about 100 nanometres. The authors made devices from three types of tungsten disulfide: a monolayer, a bilayer and a nanotube (see Fig. 1a of the paper¹).

The authors found that, whereas the monolayer and bilayer devices generated a negligible electric current under illumination, the nanotube device exhibited a large photovoltaic effect. Given that these three types of solar cell have the same, uniform chemical identity, how does the nanotube device convert light into electricity

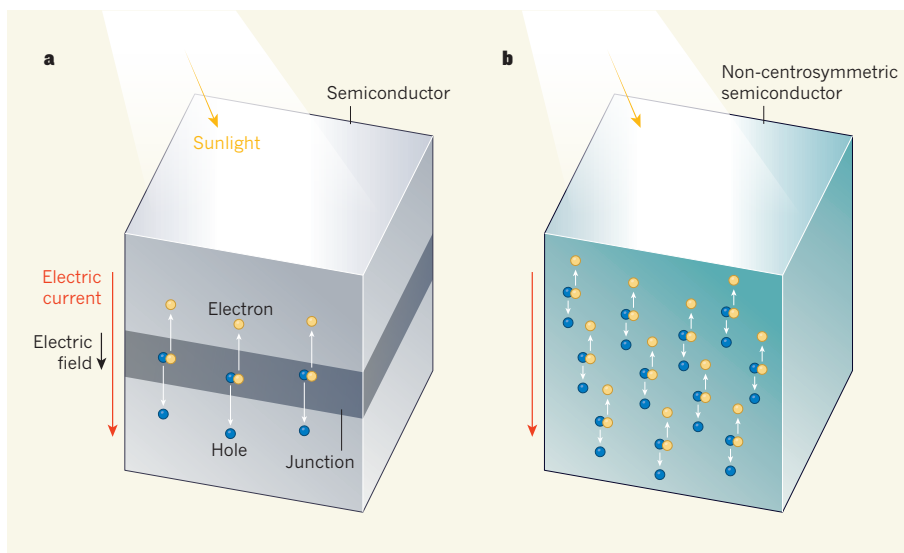


Figure 1 | Two types of solar cell. **a**, A conventional solar cell is made of a semiconductor such as silicon. Electrical transport occurs through electron vacancies called holes in one region (bottom), and through electrons in another region (top). An electric field is generated across the junction between these two regions. When this junction is illuminated by sunlight, electron-hole pairs are produced. The electrons and holes are then separated by the electric field, giving rise to an electric current. **b**, Zhang *et al.*¹ report a junction-free solar cell that is made of a non-centrosymmetric semiconductor — one whose structure lacks symmetry under a transformation known as spatial inversion. Under illumination, electron-hole pairs are produced and separated because of a phenomenon called the bulk photovoltaic effect, generating an electric current.

without the assistance of a junction? And why is this ability absent in the monolayer and bilayer devices? Zhang *et al.* point to a phenomenon called the bulk photovoltaic effect (BPVE), and attribute the diverse performance of the solar cells to their distinctive crystal symmetries. The BPVE can spontaneously generate a current in a uniform semiconductor, without the requirement for a junction (Fig. 1b).

The BPVE was first observed² at Bell Labs in New Jersey in 1956, just two years after the invention of modern silicon solar cells³. The effect is restricted to non-centrosymmetric materials⁴, which are characterized by a lack of symmetry under spatial inversion (the combination of a 180° rotation and a reflection). The BPVE has two intriguing features: the light-generated current depends on the polarization of the incident light, and the associated voltage is larger than the material's bandgap (the energy required to excite conducting free electrons). However, the effect has a typically low conversion efficiency, and so has remained of

academic rather than practical interest over the years.

To achieve a high BPVE efficiency, a material must have high light absorption and low internal symmetry. However, these two properties do not usually coexist in a given material. Semiconductors that absorb most incident sunlight generally have high symmetry, which diminishes or even prevents the BPVE. And common materials that have low symmetry, such as compounds called perovskite oxides, absorb little sunlight owing to their large bandgap. To circumvent this problem, tremendous effort has been devoted to enhancing light absorption in materials that have low symmetry, for example by using doping⁵. Meanwhile, it has been shown that the BPVE can be enabled in semiconductors when it is otherwise prohibited, by using mechanical fields to tailor the material's crystal symmetry⁶.

Zhang and colleagues' work calls attention to a hitherto-unexplored approach: shaping the semiconductors that have high light

absorption into nanotubes. In the case of tungsten disulfide, the crystal symmetry of the nanotube is reduced with respect to that of the monolayer and bilayer, because of the tube's curved walls. The combination of excellent light absorption and low crystal symmetry means that the nanotube exhibits a substantial BPVE. The density of the electric current associated with the BPVE surpasses that of the materials that have inherently low symmetry, even though the conversion efficiency of the BPVE is still much lower than that of the junction-based photovoltaic effect in conventional solar cells.

The authors' results demonstrate the great potential of nanotubes in harvesting solar energy, and raise several technological challenges and scientific questions. From

an applications perspective, it would be instructive to fabricate and characterize a solar cell that is made up of an array of semiconductor nanotubes, to check the feasibility of scaling up the approach. The direction of the current generated by the BPVE in each nanotube would be largely determined by the material's internal symmetry. Therefore, a uniform symmetry across the nanotube array would be needed to gather a collective current from the solar cell. In a worst-case scenario, if the currents generated in different nanotubes were in opposite directions, they would cancel each other out.

An important but unanswered question is whether the BPVE and the junction-based photovoltaic effect could cooperate in the same solar cell, to boost the overall efficiency. These two effects could harness solar energy in

a successive manner. Nevertheless, despite the remaining challenges, Zhang and colleagues' work provides a possible route towards the design of highly efficient, unconventional solar cells. ■

Ming-Min Yang and Marin Alexe are in the Department of Physics, University of Warwick, Coventry CV4 7AL, UK.

e-mails: mingmin.yang.1@warwick.ac.uk; m.alex@warwick.ac.uk

1. Zhang, Y. J. *et al. Nature* **570**, 349–353 (2019).
2. Chynoweth, A. G. *Phys. Rev.* **102**, 705–714 (1956).
3. Chaplin, D. M., Fuller, C. S. & Pearson, G. L. *J. Appl. Phys.* **25**, 676–677 (1954).
4. Fridkin, V. M. *Photoferroelectrics* (Springer, 1979).
5. Grinberg, I. *et al. Nature* **503**, 509–512 (2013).
6. Yang, M. M., Kim, D. J. & Alexe, M. *Science* **360**, 904–907 (2018).

EVOLUTION

Sex chromosomes manipulate mate choice

Female mate choice in some species selects for traits that are harmful to males. A hypothesis to explain how such mating preferences might evolve puts the spotlight on sex chromosomes. [SEE LETTER P.376](#)

MARK KIRKPATRICK

Among some of the most spectacular sights in nature are the bizarre mating displays of certain animals. These displays risk decreasing male survival because the bright colours and loud calls of the males might attract predators (Fig. 1). Yet despite this risk of harm, these traits have nevertheless evolved because only the males that can make the most exaggerated displays are chosen by females as mates. An enduring mystery in evolutionary biology is why females of these species have evolved such mating preferences. On page 376, Muralidhar¹ uses a population-genetics model to show how these types of preference can evolve if the versions of genes that contribute to the male mating display have beneficial effects when those versions of genes are present in females.

A variety of hypotheses for how female mating preferences evolve have been inspired by results from fieldwork, experiments and mathematical modelling. Most of these hypotheses fall into one of two main camps². In the first, mating preferences evolve by direct selection — the genes that affect mating preferences are themselves direct targets of selection. For example, if some males offer superior parental care, female mating preferences for choosing those males will spread by natural selection. This is why females in monogamous species often prefer males that

are not brightly coloured and that are therefore more likely to evade predators and survive to help rear offspring.

Direct selection can also act when genes

that affect mating preferences have a range of effects on other traits. This is the most probable explanation for why females of some species have pre-existing preferences for traits that don't exist in males of their own species. For example, males of some species of sword-tailed fishes have tails with greatly elongated lower fin rays. Females from a related species in which males do not have this 'sword' structure nevertheless have a preference for it, as assessed by their attraction to males of their own species that have had an artificial sword surgically attached to them³.

The second kind of hypothesis rests instead on indirect selection. Genetic variation in genes that affect mating preferences becomes correlated with genetic variation for other traits, and selection acting directly on the latter



Figure 1 | A colourful mating display. A vibrantly coloured male flame bowerbird (*Sericulus aureus*) engages in a mating display as a female of the species looks on. The male's bright colours make the bird easy for predators to spot, raising the conundrum of how female mating preferences evolve for traits that are harmful to males. Muralidhar¹ puts forward a hypothesis for how such mating preferences might arise. Whether this idea could explain the bowerbird case is unknown.

can cause mating preferences to evolve as a side effect. One example of this is what is called the good-genes hypothesis. In this model, for example, versions of genes that boost the immune system lead to better health, which enables the males that carry these versions to be more brightly coloured. The females that have mating preferences for the brightest males will pass these good genes, which improve survival, to their offspring, along with the versions of genes for a mating preference for bright males. As selection causes these beneficial immune-system genes to spread, they also drag along the genes that drive the preference for brighter males. This hypothesis is popular among behavioural ecologists, but the evidence in support of it is mixed⁴.

The hypothesis put forward by Muralidhar to explain the evolution of female mating preferences is also based on indirect selection. This model hinges on sexually antagonistic selection — the situation in which a version of a gene that is beneficial when present in one sex is harmful when present in the other sex. Among evolutionary geneticists, there is a growing appreciation of the prevalence of sexually antagonistic selection and its myriad consequences⁵. Imagine a genetic mutation that causes an increase in the redness of both male and female plumage. This mutation might increase male fitness by increasing mating success, but it would decrease female fitness by increasing predation. Therefore, genes that cause females to prefer to mate with redder males will cause their sons from such matings to have high fitness, whereas their daughters would have low fitness. Those two opposing effects on fitness at the population level would offset each other, and, on balance, mating preferences for redder males would be selected neither for nor against.

But could there be an exception to this logic if genes that affect mating preference are inherited in an unusual way? Consider what might happen if such genes are on sex chromosomes.

In mammals and fruit flies, males have two different sex chromosomes (X and Y), whereas females have two X chromosomes. Organisms such as birds and butterflies⁶ have sex chromosomes called Z and W, and the opposite arrangement exists — the males have two of the same type of sex chromosome (Z chromosome), whereas the females have one Z and one W chromosome. A gene that affects mating preference and is located on a W chromosome exists only in females, and never in males. Therefore, a female that has a version of a gene on the W chromosome that causes a mating preference for a trait that decreases male survival will pass that version only to her daughters. If those daughters also receive from their fathers the versions of genes for a trait that is beneficial to females, the daughters will have high fitness, and this W chromosome will spread. These nefarious chromosomes can be described as selfish — they spread regardless of their effect on male fitness.

Muralidhar's mathematical analysis shows

that genes affecting mating preference that are carried by other types of sex chromosome (X, Y and Z) also have evolutionary dynamics that differ from those of genes that affect mating preference but are carried on non-sex chromosomes. But of the four types of sex chromosome, Muralidhar found that it is the W chromosome that has the greatest potential to favour the spread of versions of genes that increase female mating preferences for exaggerated (and harmful) male traits.

Models show what is possible, but only data can reveal whether the possible is a reality. Is there evidence that sex chromosomes harbour genes that affect mating preferences? Muralidhar reviewed the genetics of mating preferences that have been reported for 36 species. In more than half of these species, there is indeed evidence that genes that affect mating preferences are carried on the sex chromosomes. Disappointingly, none of these preferences is linked to the W chromosome, but the number of existing studies available for this analysis is small.

A second opportunity to bring data to bear on this model relates to its prediction that species that have ZW chromosomes should be more prone to evolve female mating preferences for sexually antagonistic traits than those with XY chromosomes. Indeed, some researchers have concluded that species that have ZW sex chromosomes tend to have exaggerated sexual displays more often than do those with XY chromosomes⁷. Muralidhar's work provides a call for more comparative data on the inheritance of mating preferences and the connections between sex-determination systems and sexual ornamentation.

Could this selfish-sex-chromosome

hypothesis explain female mating preferences for vibrant colours, as is the case for the brilliantly coloured male flame bowerbirds (*Sericulus aureus*), which are preferred by the dull-coloured females of that species (Fig. 1)? Perhaps not. Genes that affect colour (and most other traits) tend to have similar effects on both sexes. Thus, females that choose dull-coloured males will have dull-coloured daughters that will survive well. If genes that affect mating preferences are carried on the W chromosome, this would favour the evolution of preferences for dull males.

It seems that some other explanation might be needed for cases such as that of the flame bowerbird. It could well turn out that preferences for different types of male trait evolve by different evolutionary pathways. If so, the decades-old debates over which hypotheses best explain how mating preferences evolve² might ultimately transform into discussions of which mechanisms operate most commonly in certain contexts. ■

Mark Kirkpatrick is in the Department of Integrative Biology, University of Texas, Austin, Texas 78712, USA.
e-mail: kirkp@mail.utexas.edu

1. Muralidhar, P. *Nature* **570**, 376–379 (2019).
2. Kirkpatrick, M. & Ryan, M. J. *Nature* **350**, 33–38 (1991).
3. Basolo, A. L. *Proc. R. Soc. B* **259**, 307–311 (1995).
4. Prokop, Z. M., Michalczyk, Ł., Drobniak, S. M., Herdegen, M. & Radwan, J. *Evolution* **66**, 2665–2673 (2012).
5. Rowe, L., Chenoweth, S. F. & Agrawal, A. F. *Am. Nat.* **192**, 274–286 (2018).
6. Bachtrog, D. et al. *PLoS Biol.* **12**, e1001899 (2014).
7. Reeve, H. K. & Pfennig, D. W. *Proc. Natl Acad. Sci. USA* **100**, 1089–1094 (2003).

This article was published online on 5 June 2019.

MATERIALS SCIENCE

Crazy colour

The formation of microscopic pores and fibrils in polymers under stress — a process called crazing — often preludes material failure. Controlled crazing has now been used to produce an array of colours in polymer films. [SEE LETTER P.363](#)

SEUNG HWAN KO

When a typical transparent, glassy polymer is bent or stretched, partial whitening of the material often occurs just before it cracks or fractures¹. This unpredictable phenomenon is called crazing, and has generally been seen as something to be avoided. But on page 363, Ito *et al.*² report that crazing can be fully controlled, and can be used to endow transparent polymers with colour. Controlled crazing could therefore be developed as the basis of an inkless, high-resolution method for printing colour on various flexible and transparent polymer materials.

Transparent polymers have conventionally been colored by mixing them with pigments, or by printing pigment-containing ink on polymer surfaces. However, transparent polymers can also be colored by producing microscopic structures within the materials — an effect known as structural coloration. Structural colours are frequently observed in nature, for example in butterfly wings³. Ito and co-workers use crazing as the basis for structural colour.

Crazing patterns in polymers form in a direction perpendicular to the applied stress, and consist of interpenetrating, micrometre-scale voids bridged by highly oriented polymer

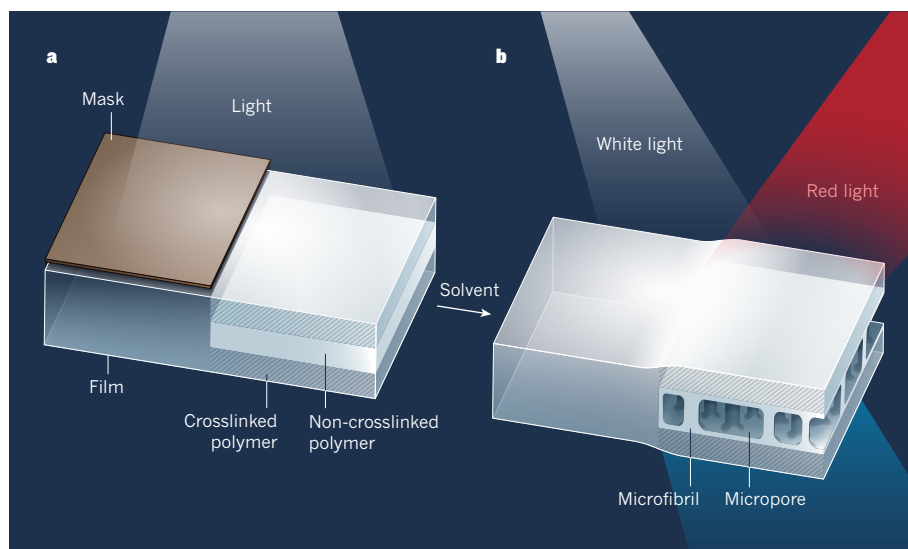


Figure 1 | Inkless colour printing in polymers. **a**, Ito *et al.*² have produced colour images in transparent polymers by shining standing waves of light through masks on polymer films. The light generates alternating layers of crosslinked and non-crosslinked polymers, which causes stress to build in the non-crosslinked layers. **b**, When treated with solvent, the non-crosslinked layers undergo crazing — microscopic pores and fibrils form, releasing the stress. This produces alternating dense and porous polymer layers, which cause the refractive index of the film to vary periodically. White light striking the layered parts of the film therefore reflects in such a way that a particular colour is produced. In principle, any colour can be generated.

microfibrils. The microvoids and microfibrils in uncontrolled crazes vary widely in size, and reflect a broad range of wavelengths of light — which explains why crazes usually look white. Ito and colleagues have demonstrated that, if crazing is controlled to generate porous layers that alternate with compact, non-porous layers, this can reinforce interference of the light reflected from the different layers, thereby producing specific colours.

The authors take advantage of a phenomenon⁴ that controls a polymer's stress field (the distribution of forces within it that balances external forces), and so controls craze generation. When a 'standing wave' light pattern is formed in a light-sensitive polymer film, crosslinks between the polymer molecules form selectively in layers, which are separated by other layers in which no crosslinking has occurred (Fig. 1); this causes tensile stress to build across the non-crosslinked layers. The authors exposed such layered films to a solvent, which releases the stress by causing crazes to form in the non-crosslinked layers. The resulting films therefore contained alternating dense and porous layers, generating periodic variations in the refractive index of the material. Light shining on the films is reflected at successive craze layers, leading to interference effects that cause structural coloration.

Ito *et al.* carried out a series of experiments to investigate the physical mechanism of, and the optimum conditions for, periodic craze formation in various transparent polymer films. The microvoids in crazes are, effectively, tiny cracks, and the authors conclude that the formation of the cracks must be controlled to control the crazing process. Their method is therefore a

real triumph: crack-formation processes are much more complex and difficult to manage in amorphous materials⁵ (such as polymer films) than in crystalline ones⁶, because the microscopic structures of amorphous materials are more random.

The authors report the production of only a few colours in their work, but a wide range should, in principle, be generated by carefully adjusting the spacing of the alternating layers. The spacing can, in turn, be controlled by

"This simple, inexpensive coloration method is based on a phenomenon that was previously regarded as useless."

altering several factors: the wavelength of light used to produce the layers and the amount of time used to irradiate the films; the type and molecular weight of the polymer; the initial thickness of the film; the type and temperature of the solvent used to produce crazing; and the period of time for which the film is immersed in the solvent.

The authors are not the first to observe this kind of structural coloration in a multi-layered transparent film. However, most research in this area has involved complex and expensive methods in which alternating layers of films are deposited on a substrate in a vacuum (see ref. 7, for example). By contrast, Ito and co-workers have developed a simple, inexpensive method based on a phenomenon that was previously regarded as useless. Indeed, previous studies of crazing have concentrated mostly on finding ways to inhibit or

prevent it^{8,9}, rather than to control it.

The authors demonstrate that craze control can be used for inkless colour printing at incredibly high resolution (up to 14,000 dots per inch). The resolution of conventional colour-printing methods, such as inkjet printing, is generally just 600–1,200 dots per inch because of limitations associated with the size of the ink droplets that can be generated and the effects of ink spreading. Another advantage of Ito and colleagues' technique is that the printing time will not depend strongly on the size of the substrate, because it is a parallel process (the whole pattern is printed into the film at the same time), whereas conventional inkjet printing is a serial writing process that takes considerably longer to print large areas.

The impact of this work is not limited to the coloration of transparent polymer materials — it will also enhance our understanding of crazing more generally. For example, the crazing described by Ito *et al.* occurs in the out-of-plane direction (the layers stack up in the direction of the film's thickness), rather than in the plane of the film, as is normally observed in polymers under stress. It is thus an intriguing phenomenon that deserves further study. It will also be interesting to explore the mechanical and electrical properties of controlled crazing.

The surprisingly simple nature of the authors' method means that it could easily be adapted for use by currently available technology for colorizing polymers. However, as with any new technology, several hurdles will need to be overcome. The authors' process is largely limited to a narrow set of operating conditions and certain material combinations, so its general applicability to other materials remains to be seen. Further work exploring the physical mechanism involved in detail might reveal how the method could be applied to any polymer material. In the meantime, craze control will probably find exciting applications beyond inkless colour printing in a transparent polymer, such as in electronic devices and sensors. ■

Seung Hwan Ko is in the Department of Mechanical Engineering, Seoul National University, Gwanak-gu, Seoul 08826, South Korea.

e-mail: maxko@snu.ac.kr

- Swallowe, G. M. *Mechanical Properties and Testing of Polymers* (Springer, 2010).
- Ito, M. M. *et al.* *Nature* **570**, 363–367 (2019).
- Vukusic, P., Sambles, J. R., Lawrence, C. R. & Wootton, R. J. *Proc. R. Soc. Lond. B* **266**, 1403–1411 (1999).
- Henderson, C. in *Encyclopedia of Microfluidics and Nanofluidics* (ed. Li, D.) 2073–2079 (Springer, 2008).
- Kim, M., Ha, H. & Kim, T. *Nature Commun.* **6**, 6247 (2015).
- Nam, K. H., Park, I. H. & Ko, S. H. *Nature* **485**, 221–224 (2012).
- Kolle, M. *et al.* *Nature Nanotechnol.* **5**, 511–515 (2010).
- Kim, S. Y., Kim, S. H., Pak, S. Y. & Youn, J. R. *J. Appl. Polym. Sci.* **125**, 3029–3037 (2012).
- Tervoort, T. A. & Govaert, L. E. *J. Polym. Sci. B* **42**, 2066–2073 (2004).

BIOCHEMISTRY

Enzymes that detoxify marine toxins

Potent microbial toxins found in shellfish are possible starting points for drug discovery, but analogues are needed for biological testing. Toxin-modification enzymes now suggest a new approach for producing these analogues.

MONICA E. MCCALLUM & EMILY P. BALSUS

There is an acute need for new medications to treat pain. Important sources of therapeutics for pain management and other human conditions are natural products — complex, biologically active small molecules made by living organisms. But compounds isolated directly from natural sources often do not have the optimal properties to be drugs. Therefore, a major challenge faced by those using natural products as leads for drug discovery is how to access a diverse range of closely related molecular structures for biological testing. This can be accomplished using chemical synthesis, but the complex structures of natural products often make that approach challenging.

The difficulties of accessing structural analogues have hampered efforts to investigate a family of natural products called paralytic shellfish toxins (PSTs) as candidate therapeutics for pain¹. Many PSTs are highly potent (they elicit a strong response from their molecular biological targets) and are therefore highly toxic, which has hindered their development as drugs and has generated interest in accessing less potent analogues. Writing in *ACS Chemical Biology*, Lukowski *et al.*² report the biosynthetic pathway that generates PSTs to which sulfo groups (SO_3^-) have been added, which are less toxic members of this family of compounds. The sulfotransferase enzymes characterized in the study modify extremely complex substrate molecules, and therefore might facilitate access to other less toxic analogues of PSTs for drug development.

PSTs are produced by marine microorganisms, including cyanobacteria and dinoflagellates^{3,4}. They are responsible for the numbness, tingling and more-severe symptoms of paralytic shellfish poisoning (caused by eating shellfish contaminated with these toxins), and interfere with the voltage-gated sodium channels that are responsible for transmitting signals in the nervous system. Previous efforts to isolate PSTs revealed that microbes often make analogues that bear one or more sulfo groups, leading to the discovery that this chemical modification reduces the potency and toxicity of these natural products^{5,6}.

The biosynthetic pathways and enzymes involved in the installation of these sulfo groups were not understood until a few years ago. The first insights were obtained from assays that used poorly characterized enzyme preparations isolated from dinoflagellates^{7,8}. These studies suggested that the sulfo groups were probably added to PSTs at a late stage of the biosynthetic pathway. More recently, the identification of the cyanobacterial genes encoding the biosynthetic machinery that produces saxitoxin, a highly potent PST, have enabled a molecular understanding of PST assembly⁹.

Saxitoxin is assembled through transformations that convert the amino acid L-arginine into a series of increasingly elaborate structures. Previous work¹⁰ had identified two putative sulfotransferase enzymes (SxtN and SxtSUL) encoded by saxitoxin's biosynthetic gene

clusters, and had found that SxtN can attach a sulfo group to a particular nitrogen atom in saxitoxin to generate an analogue called gonyautoxin 5 (Fig. 1). However, the position on saxitoxin at which the second sulfotransferase (SxtSUL) installs a sulfate (SO_4^-), and the order in which the enzymes are used in nature, were not determined.

Lukowski *et al.* have now characterized the ability of purified SxtN and SxtSUL to modify saxitoxin and other PSTs. Researchers from the same group had previously shown¹¹ that an oxygenase enzyme called GxtA catalyses the selective addition of a hydroxyl (OH) group to a normally unreactive carbon centre in saxitoxin (Fig. 1). In the current work, the authors combined SxtN and SxtSUL with GxtA, and thereby not only confirmed that SxtN installs a sulfo group on the previously identified nitrogen atom, but also discovered that SxtSUL selectively sulfates the hydroxyl group generated by GxtA.

Unlike the non-enzymatic transformations typically used in the chemical synthesis of natural products, these enzymatic reactions are highly selective for single sites on the PST scaffold, and tolerate the presence of the many densely packed, reactive chemical groups that are embedded in the complex molecular architecture of PSTs. Lukowski *et al.* were therefore able to produce a variety of sulfated PSTs directly from saxitoxin. When they measured the biological activity

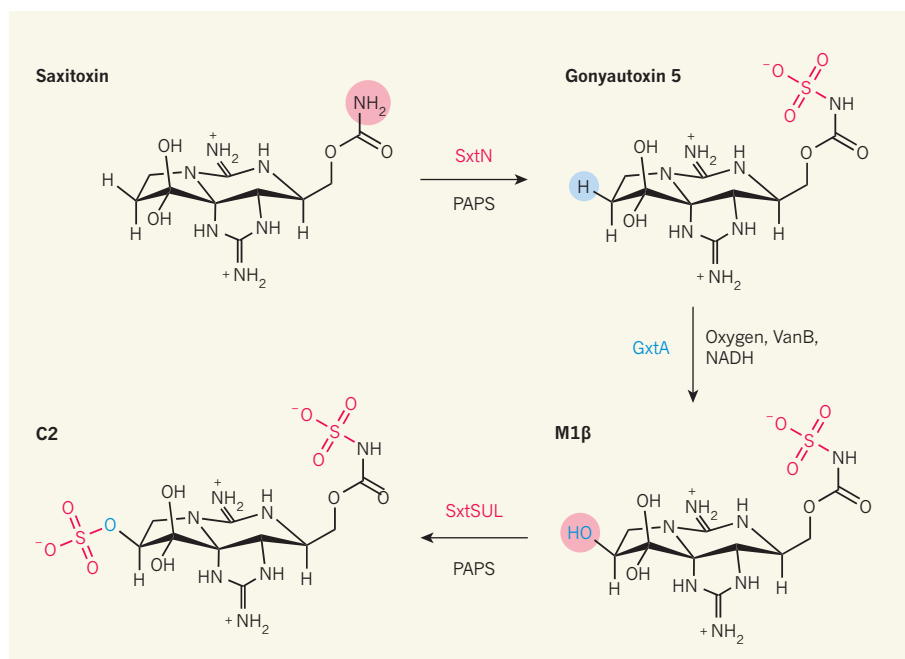


Figure 1 | Biosynthesis of sulfated paralytic shellfish toxins (PSTs). PSTs, including saxitoxin, are potentially fatal to humans, but less toxic analogues are potential leads in the search for new painkillers. Lukowski *et al.*² have worked out the biosynthetic pathway that converts saxitoxin into less toxic sulfated analogues in microbes. They find that SxtN, a sulfotransferase enzyme, attaches a sulfo group (SO_3^-) specifically to a nitrogen atom in saxitoxin, forming the compound gonyautoxin 5. The GxtA enzyme then selectively adds a hydroxyl (OH) group to the other end of the molecule, forming M1β, and a second sulfotransferase, SxtSUL, converts the hydroxyl group into a sulfate group (SO_4^-), forming the C2 analogue. The work might allow less toxic sulfated PST analogues to be prepared using a combination of conventional chemical synthesis and enzymatic chemistry. PAPS and NADH are enzyme cofactors; VanB is a partner enzyme of GxtA.

of these compounds, the results confirmed that the addition of multiple sulfo groups to PSTs reduces the compounds' binding affinities to voltage-gated sodium channels. This strongly suggests that sulfo groups reduce PST toxicity, further highlighting their potential for incorporation into PST-based drug candidates.

The use of biosynthetic enzymes to modify PSTs represents a strategy that is distinct from the chemical-synthesis approaches more frequently used to make analogues of these natural products¹². Although many of those synthetic efforts have been successful, they often involve long sequences of reactions and deliver low yields of products as a consequence of the challenging architectures of the PSTs — which contain an abundance of reactive oxygen and nitrogen atoms that complicate the use of more-standard chemical reactions. Lukowski and colleagues' findings now offer researchers the opportunity to combine conventional synthetic chemistry with biocatalysis, using enzymes to further modify PST scaffolds obtained by synthetic routes. This could potentially streamline access to sulfated versions of these natural products. It might eventually even be possible

to use this approach to make non-natural PST analogues for evaluation as candidate therapeutics.

However, substantial barriers must be surmounted before these sulfotransferase enzymes can be fully integrated into PST syntheses. Their catalytic efficiency is very low, and they have not yet been used on a large scale — Lukowski and colleagues worked at a sub-milligram scale, but multi-gram quantities of PST analogues would eventually be needed for the pre-clinical development of drug candidates. Also, the reactivity of the enzymes towards non-natural PST scaffolds, or towards members of related toxin families, has yet to be evaluated. If the reactivity and selectivity of the sulfotransferases can be optimized using enzyme engineering, these biocatalysts will become powerful synthetic tools in the search for new pain therapeutics. ■

“These findings offer researchers the opportunity to combine conventional synthetic chemistry with biocatalysis.”

Monica E. McCallum and Emily P. Balskus are in the Department of Chemistry and Chemical Biology, Harvard University, Cambridge, Massachusetts 02138, USA. e-mail: balskus@chemistry.harvard.edu

1. Durán-Riveroll, L. M. & Cembella, A. D. *Mar. Drugs* **15**, 303 (2017).
2. Lukowski, A. L. *et al.* *ACS Chem. Biol.* **14**, 941–948 (2019).
3. van Apeldoorn, M. E., van Egmond, H. P., Speijers, G. J. A. & Bakker, G. J. I. *Mol. Nutr. Food Res.* **51**, 7–60 (2007).
4. Bustillos-Guzmán, J. J. *et al.* *Food Addit. Contam. A* **32**, 381–394 (2015).
5. Andrinolo, D., Michea, L. F. & Lagos, N. *Toxicol.* **37**, 447–464 (1999).
6. Andrinolo, D., Iglesias, V., Garcia, C. & Lagos, N. *Toxicol.* **40**, 699–709 (2002).
7. Sako, Y. *et al.* *J. Phycol.* **37**, 1044–1051 (2001).
8. Yoshida, T. *et al.* *Fish. Sci.* **68**, 634–642 (2002).
9. Wang, D.-Z., Zhang, S.-F., Zhang, Y. & Lin, L. *J. Proteom.* **135**, 132–140 (2016).
10. Cullen, A. *et al.* *ACS Chem. Biol.* **13**, 3107–3114 (2018).
11. Lukowski, A. L. *et al.* *J. Am. Chem. Soc.* **140**, 11863–11869 (2018).
12. Berlinck, R. G. S., Bertonha, A. F., Takaki, M. & Rodriguez, J. P. G. *Nat. Prod. Rep.* **34**, 1264–1301 (2017).

This article was published online on 10 June 2019.

IMMUNOLOGY

A licence to kill during inflammation

Inflammasomes are protein complexes that fight infection by driving inflammation or cell death. It now seems that the protein NEK7 provides a ‘licence’ for the formation of inflammasomes containing the protein NLRP3. SEE ARTICLE P.338

KENGO NOZAKI & EDWARD A. MIAO

Inflammation can help to eliminate infection, but excessive inflammation can cause damage to the body. The sensor proteins that trigger an inflammatory immune response must therefore be carefully regulated. Some intracellular immune-sensor proteins detect components in a cell that become abnormal or altered during a cellular crisis. Signs of cellular crisis are sometimes produced in the absence of an infection, so mechanisms are needed to prevent the proteins from triggering an inappropriate inflammatory response. Sharif *et al.*¹ report a structural study on page 338 that investigates an immune-sensor protein called NLRP3, revealing that a protein called NEK7 acts as a ‘licence’ that enables this protein to cause inflammation.

When an immune sensor recognizes a hallmark of infection in the cytoplasm, this can activate the protein and lead to the assembly of a multiprotein complex called an inflammasome. The activation of proteins that

function downstream of an inflammasome can potently drive both inflammation and cell death². Different types of inflammasome can form depending on the sensor components involved. Certain inflammasomes respond to a highly specific trigger: for example, those in mammalian cells containing the sensor protein NLRC4 respond to the presence of the bacterial protein flagellin^{3,4}.

Proteins that are normally present in mammalian cells do not seem able to trigger the accidental formation of NLRC4-containing inflammasomes, given the lack of reports of such aberrant events. By contrast, inflammasomes that contain NLRP3 are activated when NLRP3 recognizes — by an as yet unknown mechanism — hallmarks of cellular catastrophe, such as extremely low concentrations of potassium in the cytoplasm, or signs of dysfunction in organelles called mitochondria². Such events can arise from tissue damage that is unrelated to infection, and NLRP3 activation in such cases has been implicated as a possible cause of inflammatory diseases such as atherosclerosis.

It is widely accepted that the tightly regulated formation of NLRP3-containing inflammasomes occurs in two steps. In the first step, NLRP3 is primed for action by other immune-sensor proteins called TLRs, which can detect components of microorganisms. This priming step occurs in two ways⁵: NLRP3 can undergo a modification, such as the addition of a phosphate group or the removal of an attached ubiquitin protein. Further priming is achieved by a rise in expression of the gene that encodes NLRP3, increasing the chance that NLRP3 will detect any abnormalities. The second step, activation, then results in NLRP3 proteins binding together to form part of a disc-shaped inflammasome complex that is probably similar to those of other inflammasomes containing proteins of the NLR family (which includes NLRP3 and NLRC4)^{5,6}. This activation step occurs during a cellular catastrophe, but the biochemical and structural mechanisms involved are unknown.

Researchers have long sought to determine the structure of NLRP3 as it forms an inflammasome, in the hope of gaining insights into how this protein functions. However, such efforts have been unsuccessful, perhaps because unknown protein partners that interact with NLRP3 were missing from earlier attempts.

The discovery^{7–9} that the enzyme NEK7 is essential for NLRP3 signalling provided a missing part of the puzzle. NEK7 regulates processes that occur during cell division, such as the breakdown of the nuclear-envelope structure¹⁰, so it was surprising to find that it has a separate role in inflammation. This suggested that NLRP3-containing inflammasome formation doesn't occur during cell division because NEK7

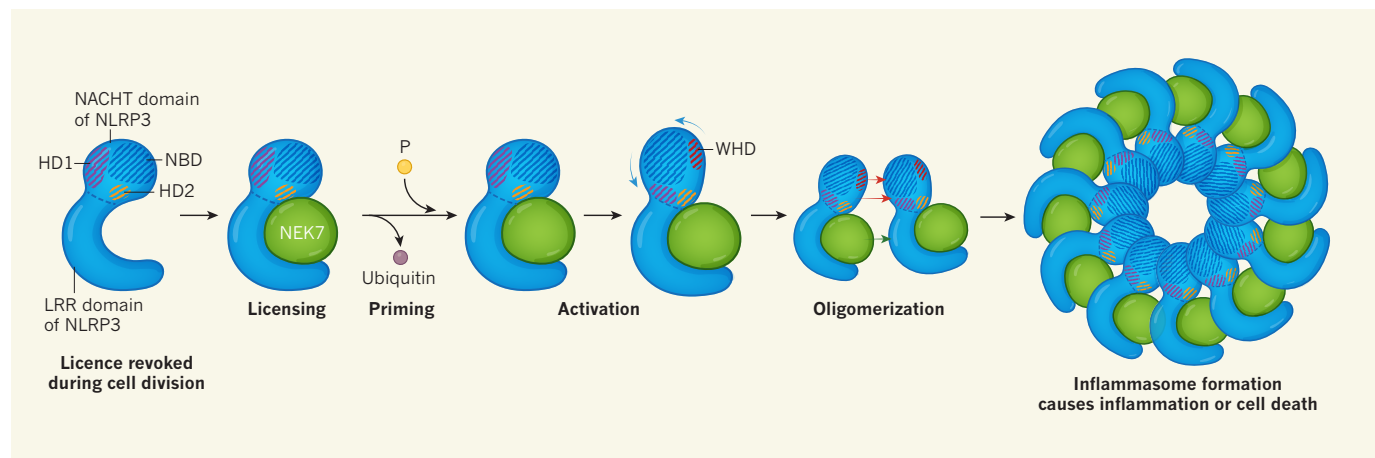


Figure 1 | The NEK7 protein enables the formation of a multiprotein immune-defence complex called the inflammasome. Sharif *et al.*¹ report structural studies of the assembly of an inflammasome containing the human protein NLRP3. Inflammasome formation requires^{7–9} the enzyme NEK7. Sharif and colleagues report that NEK7 helps the inflammasome to assemble by providing a ‘licence’ for its formation. Given that NEK7 has roles in cell division, this licensing activity is probably revoked if NEK7 is not available when cells divide. NLRP3 has a NACHT domain composed of the subdomains: NBD, HD1, WHD and HD2 (the WHD subdomain is exposed when the protein is activated) and an LRR domain. In the inactive state of NLRP3, NEK7 binds to

NLRP3’s LRR domain and to its NACHT domain in the HD2 subdomain. An early step leading to inflammasome formation is called priming, which is when NLRP3 either gains a phosphate group (P) or loses an attached ubiquitin protein². This is followed by rotational activation of the NACHT domain, which exposes all four of its subdomains. By analogy with another type of inflammasome^{5,6}, three of these subdomains (NBD, HD1 and WHD) might form interactions (red arrows) between adjacent proteins. Sharif *et al.* report that NEK7 forms a connection (green arrow) between LRR domains of adjacent NLRP3 proteins as the oligomerization of proteins occurs during inflammasome formation. The assembled inflammasome can cause inflammation or cell death.

is unavailable to aid inflammasome assembly⁸. NEK7 regulates NLRP3 signalling by binding to a region of the protein known as the LRR domain^{8,9}. However, why such an interaction between NEK7 and NLRP3 is essential for inflammasome formation has remained elusive.

To tackle this question, Sharif *et al.* used cryo-electron microscopy to investigate the structure of a human NLRP3 interacting with NEK7. The authors’ structural data reveal that NLRP3 and NEK7 bind to form a dimer in which NLRP3 is in an inactive conformation. In this state, the LRR domain of NLRP3 has a lobed, semicircular shape (Fig. 1), and the carboxy-terminal region of NEK7 nestles in the inner curve of the LRR.

NLRP3 also contains a structure called the NACHT domain, and in the inactive NLRP3–NEK7 complex, this domain is structurally very similar to the NACHT domain¹¹ of inactive NLRC4. It was previously shown^{5,6,11} that the NACHT domain of NLRC4 rotates as it transitions into an active conformation. This rotational-activation step uncovers part of the NACHT surface, enabling inflammasome formation through a protein-assembly process called oligomerization, and generating a disc-shaped NLRC4-containing inflammasome^{5,6}. The authors used this information to model a hypothetical conformation for an NLRP3-containing inflammasome.

In Sharif and colleagues’ model, the hypothetical rotational activation of NLRP3 doesn’t affect NEK7 binding, and NEK7 still fits into NLRP3’s LRR domain in the same way as in the inactive structure. Furthermore, the authors made the surprising discovery that NEK7 provides a bridge between adjacent NLRP3 proteins, by forming an interface with the LRR of the

adjacent NLRP3 in the inflammasome. The ability to form such an interface suggests that NEK7 provides a licence for NLRP3 to form part of the inflammasome. It seems that this licensing event occurs independently of both the priming and rotational-activation steps, because the authors did not include molecules that cause priming or add triggers for rotational activation in their structural studies. The results suggest a revised view of how the NLRP3-containing inflammasome is regulated, and put forward the idea that NLRP3 oligomerization requires NEK7 licensing. Taken together with evidence from earlier studies^{7–9}, it seems likely that this licence is revoked during cell division.

“The assembled inflammasome can cause inflammation or cell death.”

The function of NEK7 in the NLRP3-containing inflammasome is interesting when considered in relation to the structure of the NLRC4-containing inflammasome. The LRR domain of NLRC4 is longer than that of NLRP3, and makes direct contact with the LRR of the adjacent NLRC4 protein in the inflammasome^{5,6}. In an NLRP3-containing inflammasome, NEK7 fulfils a similar connecting role by making contact with adjacent LRR domains. This explains why NLRP3-containing inflammasomes require NEK7 licensing, whereas NLRC4-containing ones do not.

Many mysteries concerning the regulation of NLRP3-containing inflammasomes remain. Perhaps future structural studies will reveal how NLRP3 modification accomplishes the priming step.

Finally, we still don’t know the answer to

perhaps the most important question of all: what direct interaction between NLRP3 and an unknown cellular factor results in the formation of the inflammasome? Perhaps the evidence that the NEK7 licence is revoked during cell division provides a clue. If inappropriate activation of NLRP3 is likely to occur during cell division, then having an NEK7-licensing step would help to combat this potential problem. Thus, one could imagine that the type of cellular catastrophe detected by NLRP3 also occurs during cell division but in a controlled manner. ■

Kengo Nozaki and Edward A. Miao are in the Department of Microbiology and Immunology, Center for Gastrointestinal Biology and Disease, and at the Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, USA. e-mail: edmiao1@gmail.com

1. Sharif, H. *et al.* *Nature* **570**, 338–343 (2019).
2. Swanson, K. V., Deng, M. & Ting, J. P.-Y. *Nature Rev. Immunol.* <https://doi.org/10.1038/s41577-019-0165-0> (2019).
3. Miao, E. A. *et al.* *Nature Immunol.* **7**, 569–575 (2006).
4. Franchi, L. *et al.* *Nature Immunol.* **7**, 576–582 (2006).
5. Zhang, L. *et al.* *Science* **350**, 404–409 (2015).
6. Hu, Z. *et al.* *Science* **350**, 399–404 (2015).
7. Schmid-Burgk, J. L. *et al.* *J. Biol. Chem.* **291**, 103–109 (2016).
8. Shi, H. *et al.* *Nature Immunol.* **17**, 250–258 (2016).
9. He, Y., Zeng, M. Y., Yang, D., Motro, B. & Núñez, G. *Nature* **530**, 354–357 (2016).
10. Fry, A. M., Bayliss, R. & Roig, J. *Front. Cell Dev. Biol.* **5**, 102 (2017).
11. Hu, Z. *et al.* *Science* **341**, 172–175 (2013).

This article was published online on 12 June 2019.

Mitochondrial fission requires DRP1 but not dynamins

Tiago Branco Fonseca^{1,2,3,4}, Ángela Sánchez-Guerrero^{1,2,4}, Ira Milosevic^{2*} & Nuno Raimundo^{1*}

ARISING FROM J. E. Lee *Nature* <https://doi.org/10.1038/nature20555> (2016)

Mitochondrial fission is necessary for the maintenance of the mitochondrial network, and relies on the GTPase dynamin-related protein 1 (DRP1; also known as DNM1L, dynamin-1-like protein)^{1,2}. DRP1 forms helical oligomers that wrap around the mitochondrial outer membrane and scission it^{3,4}. Recently, it was proposed that DRP1 is not sufficient to execute mitochondrial fission, and that another GTPase, dynamin-2 (DNM2, also known as DYN2), was an essential component of the mitochondrial division machinery⁵. Here we report that mouse fibroblasts lacking all three mammalian dynamin proteins (DNM1, DNM2 and DNM3)⁶ (dynamin triple-knockout cells), as well as cells with knockdown of DNM2 only, do not display defects in mitochondrial fission or mitochondrial hyperfusion, which were readily detected after knockdown of DRP1, even in the dynamin triple-knockout cells. The same was observed when examining peroxisomal fission (mitochondria and peroxisomes share the fission machinery⁷). Thus, our data show that DRP1 is essential for mitochondrial and peroxisomal fission, whereas DNM1–DNM3 are dispensable.

Using wild-type and dynamin triple-knockout fibroblasts (Extended Data Fig. 1a, b; absence of DNM1–DNM3 is demonstrated in Extended Data Fig. 1c, d), we first characterized the steady-state morphology of the mitochondrial network by transfecting cells with green fluorescent protein (GFP) targeted to the mitochondrial matrix (mito-GFP) (Fig. 1a). The systematic analysis of the mitochondrial network revealed a similar number of individual mitochondria per region-of-interest (ROI) in wild-type and triple-knockout cells (Fig. 1b; defective fission would result in less individual mitochondria). The mitochondrial area (Fig. 1c) and perimeter (Fig. 1d) were similar between wild-type and triple-knockout cells, and no constricted areas were observed. To exclude interference of mito-GFP expression with mitochondria fission or fusion, we also characterized the morphology of the mitochondrial network by immunocytochemistry, using an antibody against the outer mitochondrial membrane protein TOM20 (Extended Data Fig. 1e). No significant differences were observed between wild-type and dynamin triple-knockout cells in the mitochondria number per ROI (Extended Data Fig. 1f), area (Extended Data Fig. 1g) or perimeter (Extended Data Fig. 1h). These results suggest that mitochondrial gross morphology and dynamics are independent of dynamin proteins.

Nevertheless, it was theoretically possible that the mitochondrial fission rate would be decreased in the absence of dynamin proteins, with the steady-state mitochondrial morphology remaining unaffected owing to a compensatory decrease in fusion⁸. To address this possibility, we measured the levels of several proteins involved in mitochondrial fission and fusion by western blot analysis. We found no difference in the levels of the fission protein DRP1 or mitochondrial fission factor (MFF), which recruits DRP1 to the mitochondrial outer membrane⁹ (Fig. 1e), between wild-type and triple-knockout cells. The phosphorylation of DRP1 residues Ser616 (which promotes

mitochondrial localization¹⁰) and Ser637 (which excludes DRP1 from mitochondria²) was not altered (Fig. 1e). In addition, the protein levels of the mitochondrial fusion proteins optic atrophy 1 (OPA1; both short and long isoforms), mitofusin 1 (MFN1), mitofusin 2 (MFN2), or the mitochondrial outer membrane protein voltage-dependent anion channel (VDAC; a marker of mitochondrial mass) were also indistinguishable between wild-type and dynamin triple-knockout cells (Fig. 1f).

To determine whether acute loss of DNM2, which is the main dynamin protein in non-neuronal cells, would have a different effect than the DNM1–DNM3 triple knockout, we knocked down DNM2 with approximately 95% silencing efficiency (Extended Data Fig. 1i, j), and assessed mitochondrial morphology with mito-GFP (Fig. 1g). No difference was observed in mitochondria number per ROI, area or perimeter (Fig. 1h–j) between wild-type and DNM2-knockdown cells. Similar results were also obtained by TOM20 immunostaining (Extended Data Fig. 1k–n). These data suggest that the silencing of DNM2 does not alter mitochondrial morphology in fibroblasts. The total levels of fission and fusion proteins were not affected in DNM2-knockdown cells (Extended Data Fig. 1o, p), excluding compensatory effects.

It could, however, still be argued that DNM2 has a cell-type-specific effect, not detectable in fibroblasts. Thus, we knocked down DNM2 in human HeLa cells, with over 95% efficiency (Extended Data Fig. 1q, r). No significant changes in mitochondrial morphology were detected between control (scrambled) and DNM2-knockdown HeLa cells (Extended Data Fig. 1s, t). Together, these data show that mitochondrial fission was unaffected by either the knockdown of DNM2 alone or the triple knockout of all three dynamin proteins in two different cell types.

It could be argued that because cultured cells do not present high rates of mitochondrial division or fusion under basal conditions, the removal of dynamin proteins would have no effect. We therefore tested whether silencing another protein described as essential for mitochondrial fission, DRP1, would affect the morphology of the mitochondrial network, as previously reported^{1,2}. We silenced DRP1 in wild-type and dynamin triple-knockout cells, and labelled mitochondria with mito-GFP. Despite the efficiency of DRP1 silencing being only around 70% (Extended Data Fig. 2a, b), it resulted in robust mitochondrial hyperfusion in wild-type as well as dynamin triple-knockout cells (Fig. 1k). Accordingly, DRP1 silencing resulted in an approximately 50% decrease in the number of individual mitochondria per ROI, both in wild-type and dynamin triple-knockout cells (Fig. 1l). The mitochondrial area (Fig. 1m) and perimeter (Fig. 1n) increased around twofold when DRP1 was silenced. Notable, the effect of DRP1 knockdown was similar in wild-type and dynamin triple-knockout cells (Fig. 1k–n). We obtained similar results by immunocytochemistry (Extended Data Fig. 2c–f).

¹Institute of Cellular Biochemistry, University Medical Center Göttingen, Göttingen, Germany. ²European Neuroscience Institute Göttingen, A Joint Initiative of the University Medical Center Göttingen and the Max-Planck-Society, Göttingen, Germany. ³Department of Biology, University of Padova, Padova, Italy. ⁴These authors contributed equally: Tiago Branco Fonseca, Ángela Sánchez-Guerrero. *e-mail: imilose@gwdg.de; nuno.raimundo@med.uni-goettingen.de

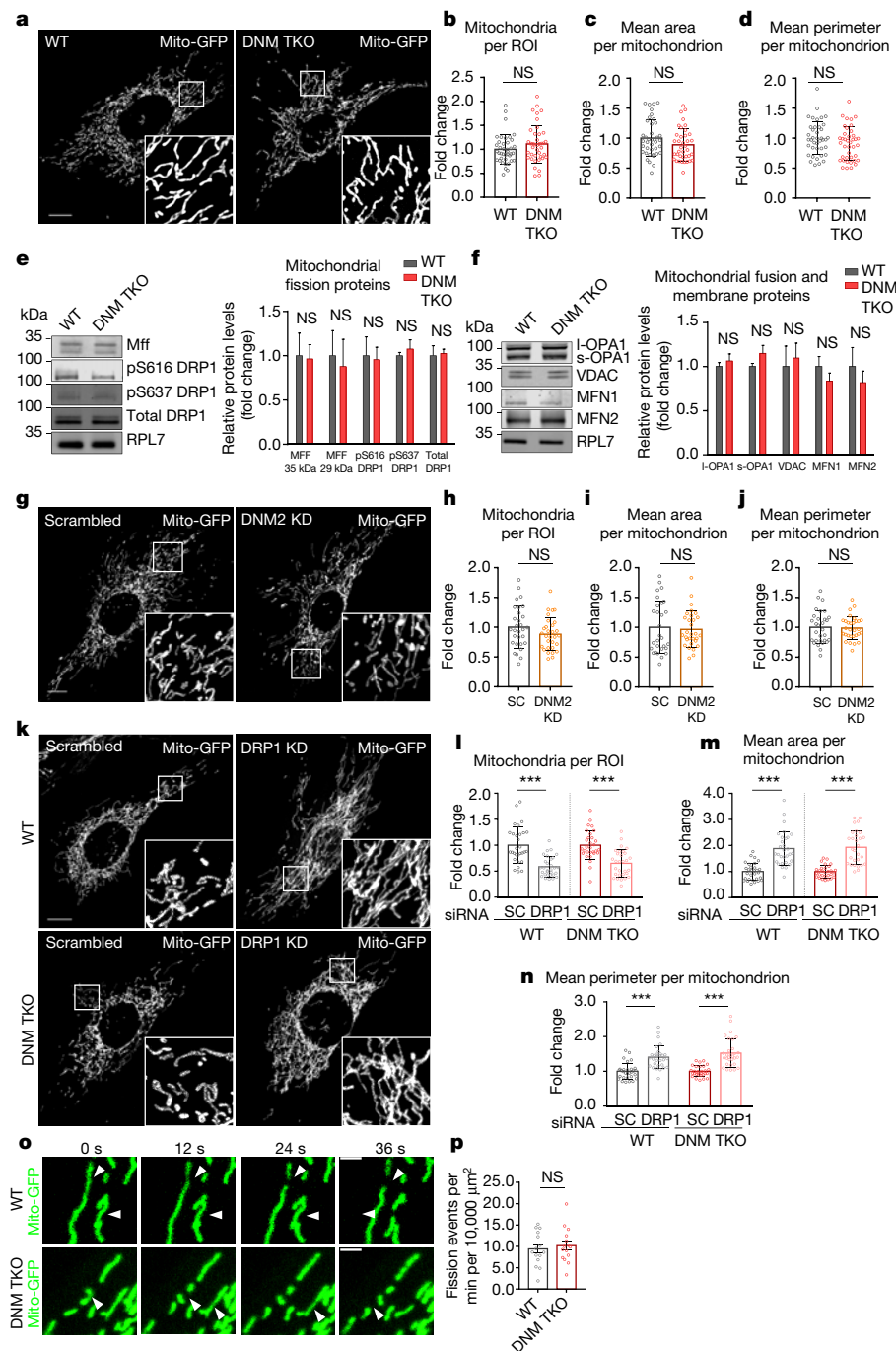


Fig. 1 | Mitochondrial morphology is not affected in the absence of DNMI1, DNMI2 and DNMI3. **a**, Wild-type (WT) and dynamin triple-knockout (DNM TKO) cells expressing Mito-GFP ($n = 40$ cells each). **b–d**, Respective mitochondrial morphology analysis of cells in **a** by number of mitochondria per ROI (**b**), and mean area (**c**) and perimeter (**d**) per mitochondrion ($n = 40$ cells each). **e, f**, Immunoblots (left) and respective quantification (right) of mitochondrial fission (**e**) and fusion (**f**) proteins ($n = 4$). RPL7 was used as a loading control. **g**, Wild-type cells transfected with scrambled or DNMI2 DsiRNA cells expressing Mito-GFP ($n = 30$ cells each). KD, knockdown. **h–j**, Mitochondrial morphology, analysed as in **b–d** ($n = 30$ cells each). **k**, Wild-type and dynamin triple-knockout cells transfected with scrambled or DRP1 DsiRNA expressing Mito-GFP ($n = 30$ cells each). **l–n**, Mitochondrial morphology, analysed as in **b–d** ($n = 30$ cells each). Data in **a–n** obtained from at least three independent experiments. **o**, Representative time-lapse frames of a mitochondrial fission event in wild-type and dynamin triple-knockout cells expressing Mito-GFP ($n = 15$ cells each). **p**, Fission events per minute and area in wild-type and dynamin triple-knockout cells ($n = 15$ cells each). Scale bars, 10 μm (**a, g, k**) and 2 μm (**o**). Error bars represent s.d. (**b–d, h–j, l–n**) and s.e.m. (**e, f, p**). *** $P < 0.001$, one-tailed Student's t -test. NS, not significant.

These results show that mitochondrial morphology is robustly affected by decreasing the levels of DRP1, but remains unchanged when the dynamin proteins are completely removed.

Next, we expressed mito-GFP in wild-type and dynamin triple-knockout mouse fibroblasts to assess mitochondrial fission dynamics by live-cell imaging: fission occurs seemingly unperturbed in triple-knockout cells (Fig. 1o; Supplementary Video 1). Accordingly, the rate of fission events is indistinguishable between wild-type and triple-knockout cells (Fig. 1p). The kinetics of mitochondrial fission were also indistinguishable between scrambled control and DNMI2-knockdown HeLa cells (Extended Data Fig. 1u; Supplementary Video 2). Therefore, the steady-state level of mitochondrial fission or fusion is unaltered in dynamin triple-knockout and DNMI2-knockdown cells, under basal culture conditions.

We then sought to determine whether dynamin proteins would be required under higher rates of mitochondrial fission. We thus treated the cells with staurosporine, which activates mitochondrial fission⁵. Mitochondria fragmented in staurosporine-treated wild-type and dynamin triple-knockout fibroblasts (Fig. 2a), with an increased number of mitochondria (Fig. 2b), and decreased area (Fig. 2c) and perimeter (Fig. 2d). Therefore, staurosporine-induced mitochondrial fission also occurs without dynamin proteins. However, when silencing DRP1, both the wild-type plus DRP1-knockdown cells and the dynamin-triple-knockout plus DRP1-knockdown cells were resistant to staurosporine-induced fission (Fig. 2e), with no changes in mitochondrial number, area or perimeter (Fig. 2f–h). These results show that staurosporine-induced fission requires DRP1, but not dynamin proteins.

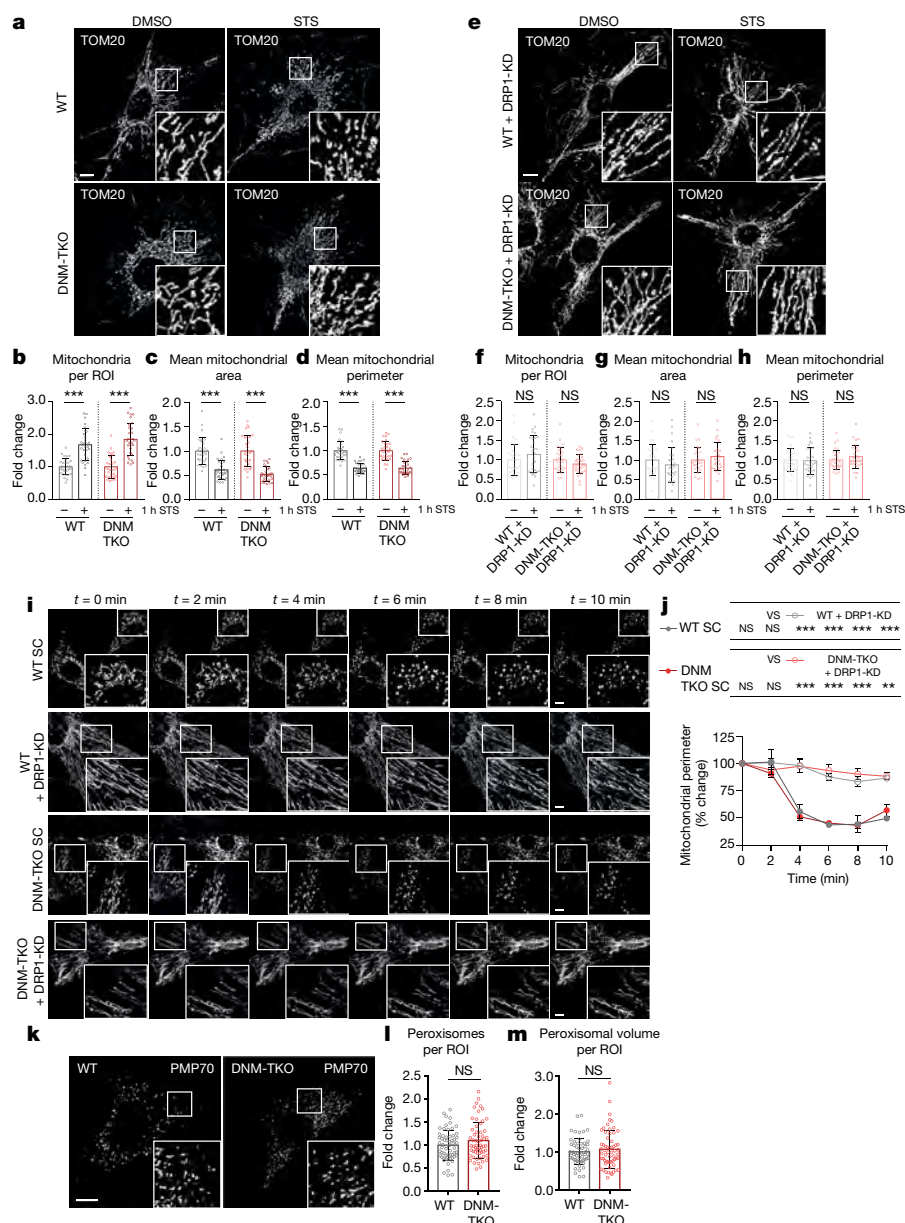


Fig. 2 | DRP1, but not dynamin proteins, is necessary for induced-mitochondrial fission.

a, TOM20 immunofluorescence of wild-type and dynamin triple-knockout cells treated with dimethylsulfoxide (DMSO) or 1 μ M staurosporine (STS) ($n = 30$ cells each). **b–d**, Respective mitochondrial morphology analysis of cells in **a**, as in Fig. 1b–d ($n = 30$ cells each). **e**, TOM20 immunofluorescence of wild-type plus DRP1-knockdown cells, and dynamin triple-knockout plus DRP1-knockdown cells, treated as in **a** ($n = 30$ cells each). **f–h**, Respective mitochondrial morphology analysis of cells in **e** ($n = 30$ cells each). **i**, Representative time-lapse frames of wild-type and dynamin triple-knockout cells transfected with scrambled (SC) or DRP1 DsiRNA and Mito-GFP exposed to 25 μ M of CCCP at $t = 2$ min. $n = 13$ cells (WT SC), $n = 17$ cells (DNM-TKO SC), $n = 7$ (WT DRP1-KD and DNM-TKO DRP1-KD). **j**, Mitochondrial perimeter analysis during CCCP treatment. $n = 13$ cells (WT SC), $n = 17$ cells (DNM-TKO SC), $n = 7$ (WT DRP1-KD and DNM-TKO DRP1-KD). Peroxisomes do not require dynamins 1–3 for fission. **k**, PMP70 immunofluorescence of wild-type and dynamin triple-knockout cells ($n = 60$ cells each). **l, m**, Peroxisomal morphology quantified by number of peroxisomes per ROI and peroxisomal volume per ROI ($n = 60$ cells each). Scale bar, 10 μ m. Data obtained from at least three independent experiments. Error bars represent s.d. (**b–d**, **f–h**, **l, m**) and s.e.m. (**j**). $**P < 0.01$, $***P < 0.001$, one-tailed Student's t -test.

We next studied mitochondrial fission kinetics using CCCP, which depolarizes the mitochondrial inner membrane and triggers mitochondrial fission¹¹. We followed the fibroblasts for two minutes before adding CCCP, and for eight minutes after by live-cell imaging. Although wild-type and dynamin triple-knockout cells (transfected with scrambled siRNA control) show pervasive mitochondrial fission shortly after the addition of CCCP, the silencing of DRP1 in both cell types confers resistance to CCCP-induced fission (Fig. 2i, Supplementary Video 3). The mitochondrial perimeter decreases prominently after CCCP addition in wild-type and in dynamin triple-knockout cells, but not in wild-type plus DRP1-knockdown cells, or in dynamin triple-knockout plus DRP1-knockdown cells (Fig. 2j). The kinetics of mitochondrial fragmentation after CCCP treatment were indistinguishable between wild-type and dynamin triple-knockout cells (Fig. 2j). These results further solidify the conclusion that mitochondrial fission, both in basal and induced conditions (for example, with staurosporine and CCCP), requires DRP1 but not dynamin proteins.

We further examined whether DNM2 is present at mitochondrial fission sites by co-transfecting mito-GFP and DNM2-mRFP in

wild-type fibroblasts and performing live imaging. DNM2-mRFP was present predominantly in the cytoplasm and in few well-defined puncta that were not at the plasma membrane. Yet it was rarely found at or near mitochondria. None of 63 unequivocally detected events of mitochondrial fission in 21 cells from three independent experiments coincided with the presence of DNM2 (Extended Data Fig. 2g, Supplementary Video 4). We next co-expressed DRP1-GFP, DYN2-mCherry (preferred here since its signal was slightly brighter than RFP) and mitochondria-targeted blue fluorescent protein (mito-BFP) in wild-type fibroblasts and HeLa cells for around 22–28 h: all discernible fission events showed DRP1-GFP on the mitochondrial fission site, in both fibroblasts (Extended Data Fig. 2h, Supplementary Video 5) and HeLa cells (Extended Data Fig. 2i, Supplementary Video 6). We could detect only 2 events in which DNM2 and DRP1 colocalized at the mitochondria fission site in 96 fission events observed in 27 cells. Notably, few DNM2 puncta not colocalized with DRP1 or fission sites were present at the mitochondria, raising the possibility that DNM2 may be involved in other processes unrelated to DRP1-dependent mitochondrial fission, such as the scission of mitochondria-derived vesicles and/or microtubule-related trafficking.

Furthermore, it has been well established that DRP1 catalyses the fission of both mitochondria and peroxisomes⁷, thus we inspected whether dynamin proteins are needed for peroxisomal fission. First, we assessed peroxisomal morphology by immunocytochemistry in wild-type and dynamin triple-knockout fibroblasts, using PMP70 as peroxisomal marker (Fig. 2k). We found no difference in peroxisome number (Fig. 2l) or volume (Fig. 2m). To eliminate the possibility that peroxisomal fission is not needed under basal culture conditions, we stimulated peroxisomal proliferation by co-expressing PEX11 β and DRP1⁷ (Extended Data Fig. 3a). As expected, the number of peroxisomes increased (Extended Data Fig. 3b, quantified in Extended Data Fig. 3c), whereas the peroxisomal area remained unchanged (Extended Data Fig. 3d). Notably, when we expressed PEX11 β and DRP1, we observed no difference between wild-type and dynamin triple-knockout cells (Extended Data Fig. 3e), both in peroxisomal number (Extended Data Fig. 3f) and area (Extended Data Fig. 3g). These results show that dynamin proteins are not needed for peroxisomal fission, either in basal or stimulated conditions. These observations were further supported by live imaging of peroxisomal fission events: fission occurred seemingly unperturbed in dynamin triple-knockout cells, with similar kinetics to wild-type cells (Extended Data Fig. 3h).

In conclusion, our results show that dynamin proteins are not required for mitochondrial fission, under both basal and stimulated conditions. However, mitochondrial fission clearly depends on DRP1, as even a partial reduction of DRP1 caused a robust hyperfusion phenotype. A similar paradigm seems to be valid for peroxisomal fission. All known essential components of mitochondrial fission and fusion (for example, DRP1, MFF, MFN1 or MFN2) present strong phenotypes in fibroblasts, pointing to fibroblasts as a good model for the studies of mitochondrial fission. In addition, unaltered mitochondrial fission rates and morphology were also observed in HeLa cells after knockdown of DNM2. This is not surprising as HeLa cells and fibroblasts have similar rates of fission–fusion events¹². Thus, the selection of cellular models is probably not an explanation for why our data are in contrast with the earlier report that used HeLa and Cos7 cells with knockdown of DNM2 as experimental model⁵. Regarding the question as to whether DRP1 can catalyse membrane scission fission itself, it was recently shown that when the outer mitochondrial membrane-specific lipid cardiolipin was present, DRP1 GTPase activity was enhanced¹³, and that DRP1 could constrict and sever membranes *in vitro*¹⁴. Although we cannot exclude the possibility that DNM2 participates in mitochondrial fission under conditions that differ from those tested here, our data unequivocally show that dynamin proteins are not essential for this process.

Methods

Dynamin-1, dynamin-2 and dynamin-3 floxed fibroblasts were prepared as previously described⁶. Control and dynamin triple-knockout cells were either electroporated with pAcGFP1-Mito (Clontech, 632432) or seeded without transfection, fixed (room temperature, 3.7% paraformaldehyde, 15 min) and stained for TOM20. Mitochondrial morphology was quantified as previously described⁵, from Z-stacks of pAcGFP1-Mito-expressing cells or non-transfected cells stained for TOM20. DRP1 or DNM2 was silenced with Dicer-substrate short interfering RNAs (DsiRNAs). Live-cell imaging was performed in a Nikon/PerkinElmer spinning-disk microscope. DMSO, staurosporine (1 μ M) and CCCP (25 μ M) were added to induce mitochondrial fission. Peroxisomal morphology was quantified in IMARIS in PMP70 immunostainings and peroxisomal fission visualized with Ub-RFP-SKL plasmid. Odyssey infrared imaging system (LICOR) was used for immunoblot acquisition. Additional information is available as Supplementary Data.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

Data supporting the findings of this study are available from the authors upon reasonable request.

Received: 20 September 2018; Accepted: 4 April 2019;

Published online 19 June 2019.

1. Bleazard, W. et al. The dynamin-related GTPase Dnm1 regulates mitochondrial fission in yeast. *Nat. Cell Biol.* **1**, 298–304 (1999).
2. Cereghetti, G. M. et al. Dephosphorylation by calcineurin regulates translocation of Drp1 to mitochondria. *Proc. Natl Acad. Sci. USA* **105**, 15803–15808 (2008).
3. Kalia, R. et al. Structural basis of mitochondrial receptor binding and constriction by DRP1. *Nature* **558**, 401–405 (2018).
4. Ji, W. K., Hatch, A. L., Merrill, R. A., Strack, S. & Higgs, H. N. Actin filaments target the oligomeric maturation of the dynamin GTPase Drp1 to mitochondrial fission sites. *eLife* **4**, e11553 (2015).
5. Lee, J. E., Westrate, L. M., Wu, H., Page, C. & Voeltz, G. K. Multiple dynamin family members collaborate to drive mitochondrial division. *Nature* **540**, 139–143 (2016).
6. Park, R. J. et al. Dynamin triple knockout cells reveal off target effects of commonly used dynamin inhibitors. *J. Cell Sci.* **126**, 5305–5312 (2013).
7. Schrader, M., Costello, J. L., Godinho, L. F., Azadi, A. S. & Islinger, M. Proliferation and fission of peroxisomes—an update. *Biochim Biophys Acta* **1863**, 971–983 (2016).
8. Sesaki, H. & Jensen, R. E. Division versus fusion: Dnm1p and Fzo1p antagonistically regulate mitochondrial shape. *J. Cell Biol.* **147**, 699–706 (1999).
9. Otera, H. et al. Mff is an essential factor for mitochondrial recruitment of Drp1 during mitochondrial fission in mammalian cells. *J. Cell Biol.* **191**, 1141–1158 (2010).
10. Kashatus, J. A. et al. Erk2 phosphorylation of Drp1 promotes mitochondrial fission and MAPK-driven tumor growth. *Mol. Cell* **57**, 537–551 (2015).
11. Ishihara, N. et al. Mitochondrial fission factor Drp1 is essential for embryonic development and synapse formation in mice. *Nat. Cell Biol.* **11**, 958–966 (2009).
12. Karbowski, M. et al. Quantitation of mitochondrial dynamics by photolabeling of individual organelles shows that mitochondrial fusion is blocked during the Bax activation phase of apoptosis. *J. Cell Biol.* **164**, 493–499 (2004).
13. Francy, C. A., Clinton, R. W., Fröhlich, C., Murphy, C. & Mears, J. A. Cryo-EM Studies of Drp1 reveal cardiolipin interactions that activate the helical oligomer. *Sci. Rep.* **7**, 10744 (2017).
14. Kamerkar, S. C., Kraus, F., Sharpe, A. J., Pucadyil, T. J. & Ryan, M. T. Dynamin-related protein 1 has membrane constricting and severing abilities sufficient for mitochondrial and peroxisomal fission. *Nat. Commun.* **9**, 5239 (2018).

Acknowledgements This work was funded by ERG StG 337327 MitoPexLysoNETWORK (N.R.), DFG E. Noether MI-1702/1 (I.M.), Schram Stiftung (I.M.) and DFG SFB1190-P02 (I.M., N.R.), ASPPA Association and Fondazione Cariparo (T.B.F.). We thank P. de Camilli for the dynamin triple-knockout cells.

Author contributions N.R. and I.M. designed and supervised the study. T.B.F., A.S.-G., I.M. and N.R. performed the experiments, analysed data and prepared the figures. N.R. wrote the manuscript which was then edited by all authors.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-019-1296-y>.

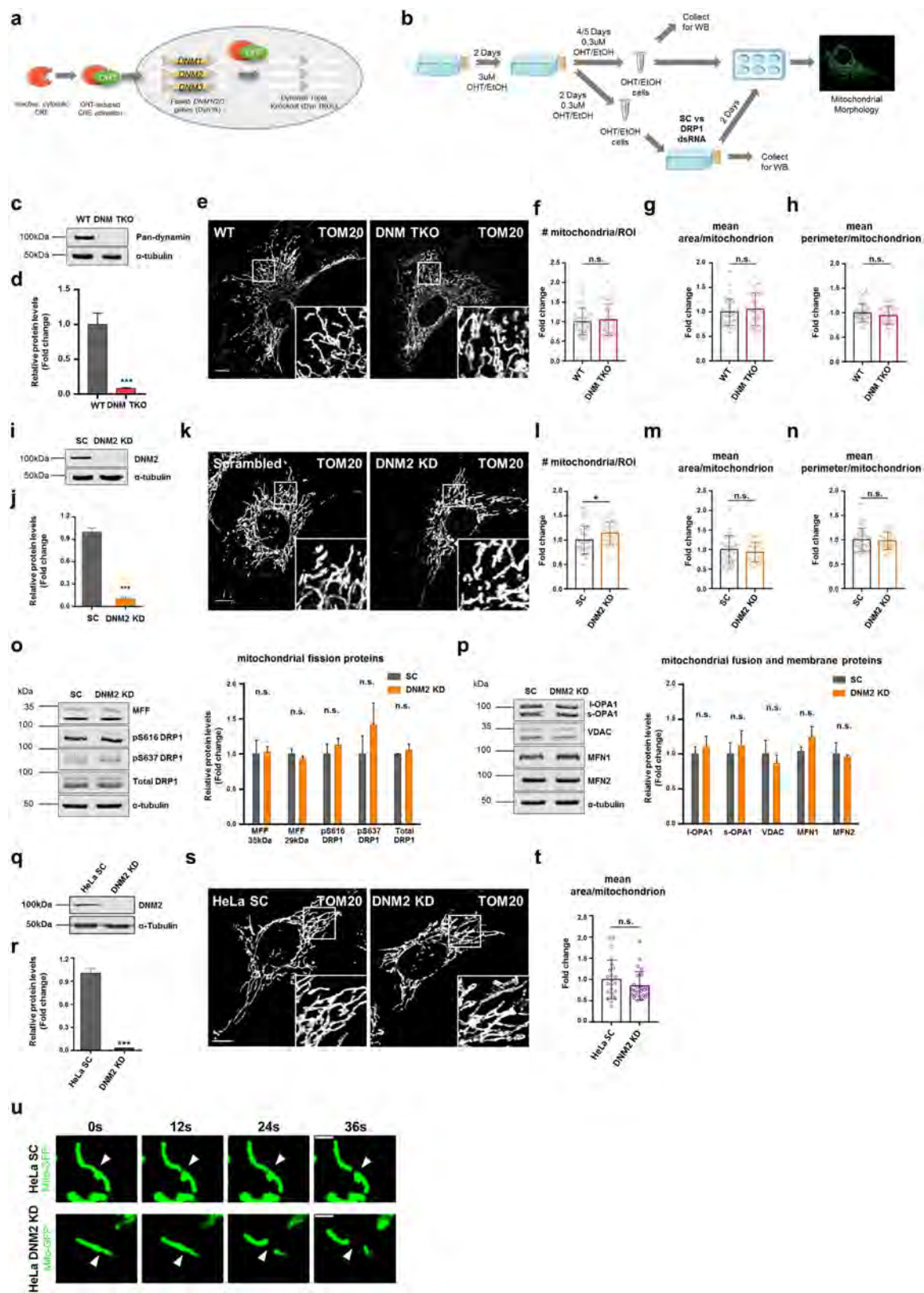
Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-019-1296-y>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to I.M. or N.R. **Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

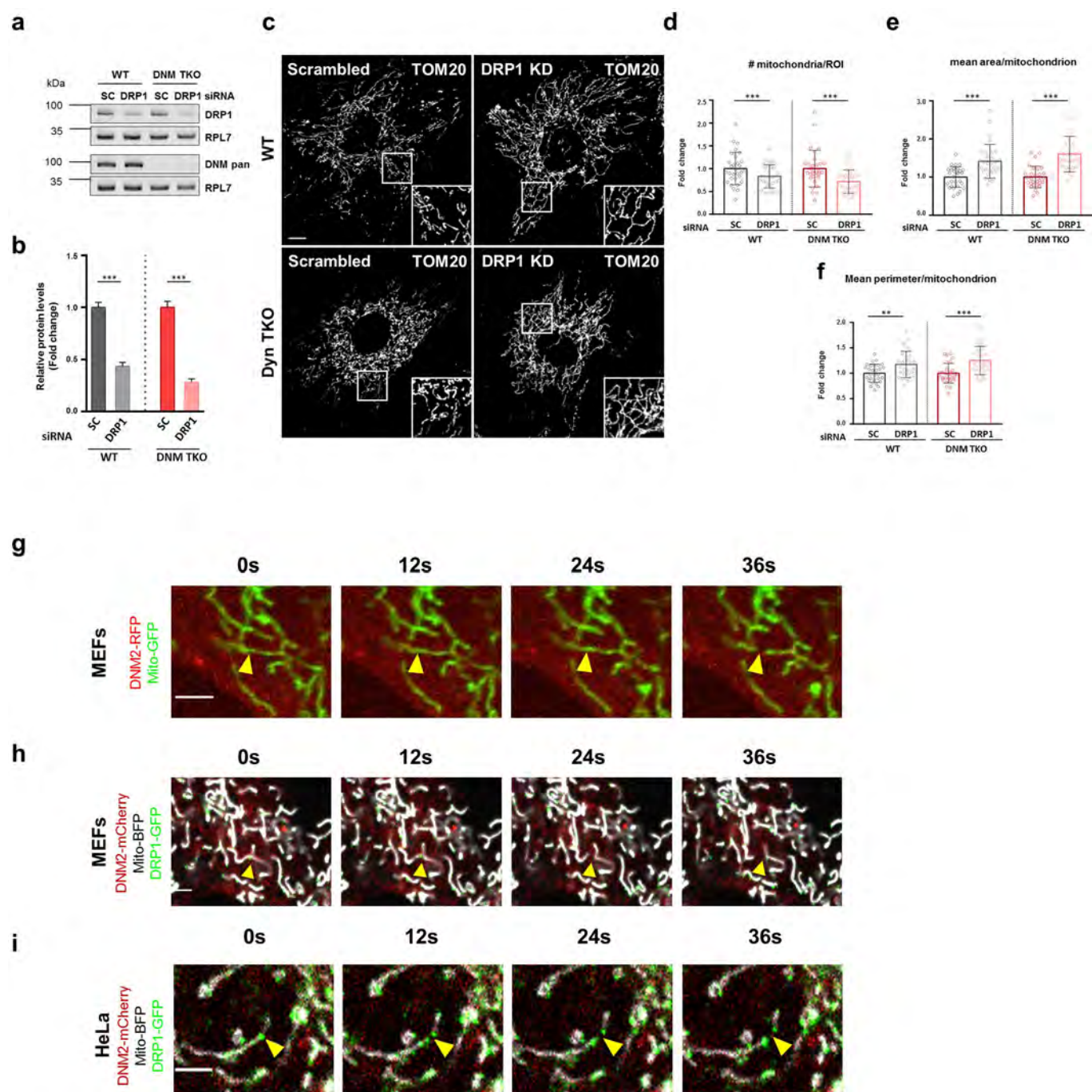
MATTERS ARISING



Extended Data Fig. 1 | See next page for caption.

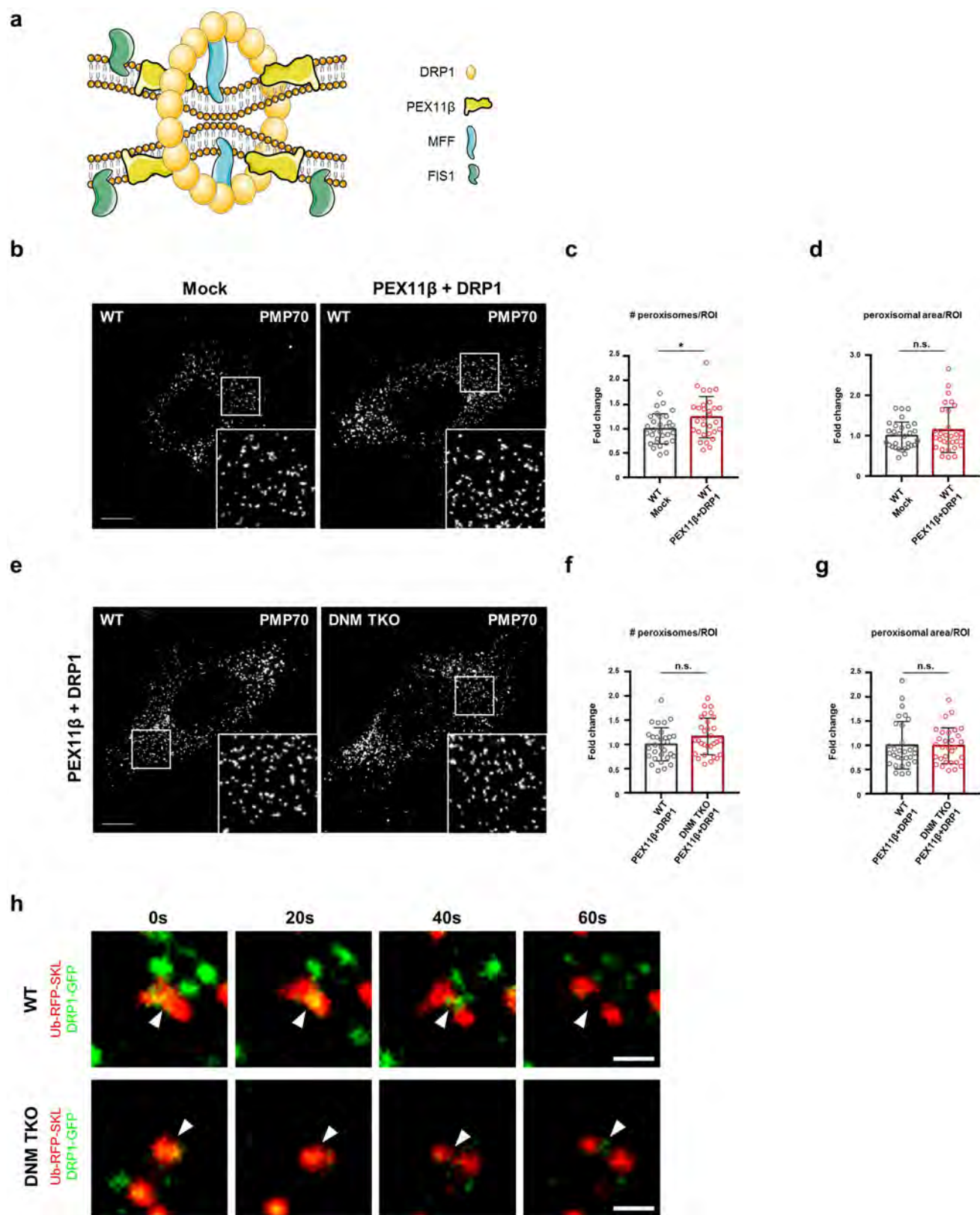
Extended Data Fig. 1 | Schematic representation of the experimental approach. **a, b**, Cre-*loxP* recombination system (**a**) and protocol to obtain dynamin triple-knockout and DRP1-knockdown cells (**b**). **c, d**, Representative immunoblot (**c**) and quantification (**d**) of expression levels of DNM1–DNM3 ($n = 3$). **e–h**, Mitochondrial fission occurs independently of DNM1, DNM2 and DNM3. **e**, TOM20 immunofluorescence of wild-type and dynamin triple-knockout cells ($n = 30$ cells each). **f–h**, Respective mitochondrial morphology analysis of the cells in **e**, as in Fig. 1b–d ($n = 30$ cells each). **i, j**, Immunoblot (**i**) and quantification (**j**) of DNM2 in scrambled and DNM2-knockdown fibroblasts ($n = 3$). α -tubulin was used as a loading control. **k**, TOM20 immunofluorescence of fibroblasts transfected with scrambled and DNM2 DsiRNA ($n = 30$ cells each). **l–n**, Respective mitochondrial

morphology analysis of cells in **k** ($n = 30$ cells each). **o, p**, Immunoblot and quantification analyses of mitochondrial fission (**o**) and fusion (**p**) proteins of DNM2 in fibroblasts transfected with scrambled and DNM2 DsiRNA ($n = 3$). **q, r**, Immunoblot (**q**) and quantification (**r**) of DNM2 in scrambled and DNM2-knockdown HeLa cells ($n = 3$). **s**, TOM20 immunofluorescence of HeLa cells transfected with scrambled and DNM2 DsiRNA ($n = 25$ cells each). **t**, Respective mitochondrial area of cells in **s** ($n = 25$ cells each). **u**, Representative time-lapse of a mitochondrial fission event in scrambled and DNM2-knockdown HeLa cells expressing Mito-GFP ($n = 15$ cells each). Error bars represent s.d. (**f–h**, **l–n**, **t**) and s.e.m. (**d**, **j**, **o**, **p**, **r**). * $P < 0.05$, *** $P < 0.001$, one-tailed Student's *t*-test. Scale bars, 10 μm (**e**, **k**, **s**) and 2 μm (**u**).



Extended Data Fig. 2 | DRP1, but not dynamin proteins, is necessary for induced-mitochondrial fission. **a**, Immunoblots of DNM1, DNM2 and DNM3 (Dyn pan) and DRP1 in scrambled and DRP1-knockdown cells in both wild-type and dynamin triple-knockout cells. **b**, Relative quantification of data in **a** ($n = 3$). **c**, TOM20 immunofluorescence of wild-type and dynamin triple-knockout cells transfected with scrambled or DRP1 DsiRNA ($n = 30$ cells each). **d–f**, Mitochondrial morphology analysis of cells in **c**, as in Fig. 1b–d ($n = 30$ cells each). **g–i**, Representative

time-lapse frames of mitochondrial fission events in wild-type mouse embryonic fibroblasts (MEFs) co-transfected with DNM2-mRFP and mito-GFP ($n = 21$ cells) (**g**), or with DRP1-GFP, DNM2-mCherry and mito-BFP ($n = 15$ cells) (**h**) and of HeLa cells expressing DRP1-GFP, DNM2-mCherry and mito-BFP ($n = 12$ cells) (**i**). Scale bars, 10 μm (**c**) and 2 μm (**g–i**). Error bars represent s.d. (**d–f**) and s.e.m. (**b**). $^{**}P < 0.01$, $^{***}P < 0.001$, one-tailed Student's t -test.



Extended Data Fig. 3 | See next page for caption.

Extended Data Fig. 3 | Peroxisomes do not require dynamin proteins for fission. **a**, Schematic model of peroxisomal fission induction by PEX11 β and DRP1 overexpression (adapted from ref. ⁷). PEX11 β oligomerizes at the peroxisome division site, promoting peroxisomal membrane elongation, enrichment of fission factors such as FIS1 and MFF, and recruitment of DRP1 at the division site for the scission event. **b**, PMP70 immunofluorescence of wild-type mock mouse embryonic fibroblasts and wild-type fibroblasts transfected with PEX11 β and DRP1 plasmids ($n = 30$ cells each). **c**, **d**, Respective quantification of peroxisomal morphology

by number of peroxisomes per ROI (**c**) and peroxisomal area per ROI (**d**) for the cells in **b** ($n = 30$ cells each). **e**, PMP70 immunofluorescence of wild-type and dynamin triple-knockout cells co-transfected with PEX11 β and DRP1 ($n = 30$ cells each). **f**, **g**, Respective peroxisomal morphology of cells in **e**, as in **c** and **d** ($n = 30$ cells each). **h**, Representative time-lapse frames of peroxisomal fission events in wild-type and dynamin triple-knockout cells, co-transfected with the peroxisomal marker Ub-RFP-SKL and DRP1-GFP ($n = 21$ cells). Scale bars, 10 μm (**b**, **e**) and 2 μm (**h**). Error bars represent s.d. * $P < 0.05$, one-tailed Student's t -test.

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a Confirmed

- ☐ ☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
- ☐ ☒ An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- ☐ ☒ The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- ☒ ☐ A description of all covariates tested
- ☐ ☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- ☐ ☒ A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- ☐ ☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted
Give P values as exact values whenever suitable.
- ☒ ☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- ☒ ☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- ☒ ☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
- ☐ ☒ Clearly defined error bars
State explicitly what error bars represent (e.g. SD, SE, CI)

Our web collection on [statistics for biologists](#) may be useful.

Software and code

Policy information about [availability of computer code](#)

Data collection

The microscope was operated by Zen software Blue 2.1. version (Zeiss). The fluorescent signal of the respective antibodies was acquired using the Odyssey infrared imaging system (LICOR).

Data analysis

- Western Blots were analyzed using ImageJ software [Schindelin, J.; Arganda-Carreras, I. & Frise, E. et al. (2012), "Fiji: an open-source platform for biological-image analysis", Nature methods 9(7): 676-682].
- Mitochondrial morphology was analysed with the Mito-Morphology macro on ImageJ [Dagda, R.K., Cherra III, S. J, Kulich, S.M. , Tandon, A , Chu, Park, D., Chu, C.T. Loss of PINK1 function promotes autophagy through effects on fission in neurons. J.Biol Chem. 284(20):13843-55, 2009].
- Data was collected and analyze with Graph Pad Prism V6 software.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/authors/policies/ReportingSummary-flat.pdf](https://www.nature.com/authors/policies/ReportingSummary-flat.pdf)

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Imaging experiments: three independent experiments were conducted with a minimum of 10 cells imaged per experiment. Wester Blot experiments: a minimum of three independent experiments were conducted with one sample per experiment.
Data exclusions	GraphPad software was used to detect outliers. Each detected outlier was excluded from the analysis.
Replication	All the experiments could be replicated at least 3 times.
Randomization	No randomization was applied because data collection and analysis was performed blindly.
Blinding	Investigators were blinded during data collection and analysis.

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
<input type="checkbox"/>	<input checked="" type="checkbox"/> Unique biological materials
<input type="checkbox"/>	<input checked="" type="checkbox"/> Antibodies
<input type="checkbox"/>	<input checked="" type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Human research participants

Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

Unique biological materials

Policy information about [availability of materials](#)

Obtaining unique materials Materials are available from standard commercial sources or from the authors.

Antibodies

Antibodies used

Tom20 (11802-1-AP, Proteintech; lot 00049639), PMP70 (SAB4200181, Merck; lot 116M4828V), HPRT (ab10479, Abcam; lot GR3226686-1), RPL7 (ab72550, Abcam; lot GR251811-11), α -Tubulin (11H10, CST; lot 9), Opa1 (BD Biosciences, 612606; lot 6224762), Mfn1 (sc-166644, Santa Cruz Biotechnology; lot GR232508-13), Mfn2 (ab56889, Abcam; lot GR219517-2), Drp1 (8570, CST; lot 2), Phospho-Ser616-Drp1 (3455, CST; lot 4), Phospho-Ser637-Drp1 (4867, CST; lot 3), Mff (17090-1-AP, Proteintech; lot 00016933), VDAC (ab110326, Abcam; lot GR130818-25), pan-dynamin (610245, BD Transduction Laboratories; lot 3242866), Dynamin-2 Ura (gift from Pietro De Camilli [Yale University, New Haven, USA]).

Validation

These antibodies have been extensively utilized in prior literature and validated by genetic approaches (knock-out, knock-down, mutation of phosphorylation sites).

Eukaryotic cell lines

Policy information about [cell lines](#)

Cell line source(s)

All the cells used on the present project are Dynamin-1, -2 and -3 floxed fibroblasts transgenic for 4-hydroxitamoxifen (OHT)-inducible Cre recombinase (Cre-ER), a kind gift from Pietro de Camilli.

The generation and characterization of these immortalized cells and the protocols for its use were first described in the following 2 papers:

- Ferguson, S.M., Raimondi, A., Paradise, S., Shen, H., Mesaki, K., Ferguson, A., Destaing, O., Ko, G., Takasaki, J., Cremona, O.,- O'Toole, E., De Camilli P. Coordinated actions of actin and BAR proteins upstream of dynamin at endocytic clathrin-coated pits. *Developmental Cell* 17, 811-822, 2009. PMID: 20059951

- Park R, Shen H, Liu L, Liu X, Ferguson S.M., De Camilli P. Dynamin triple knockout cells reveal off target effects of commonly used dynamin inhibitors. *J Cell Sci.* 2013 Sep 17. PMID: 24046449.

HeLa cells were obtained from ATCC.

Authentication

The fibroblast cells were not authenticated, but we verified that the dynamins were absent by Western blot. The HeLa cells were obtained from ATCC, and were not authenticated.

Mycoplasma contamination

The cells obtained from Pietro De Camilli's lab were tested negative for mycoplasma contamination.

Commonly misidentified lines (See [ICLAC](#) register)

The only line commonly misidentified that was used in our study is the HeLa cell line (which we obtained from ATCC).

Egg pigmentation probably has an early Archosaurian origin

Matthew D. Shawkey^{1*} & Liliana D'Alba¹

ARISING FROM J. Wiemann et al. *Nature* <https://doi.org/10.1038/s41586-018-0646-5> (2019)

Colours have crucial roles in the lives of organisms, from camouflage to mate and pollinator attraction. Although the colours of animals known only from fossils were once thought to be unknowable, recent advances have enabled some to be reconstructed, with important implications for their (Palaeo) ecology and evolution¹. A recent paper² used Raman spectroscopy to show evidence of preservation of the two colour-producing pigments in avian eggs (biliverdin and protoporphyrin IX) in fossil dinosaur eggs. The authors used these data to infer the colours of these eggs and, based on their presence and absence in extinct and extant clades, to suggest that colourful eggs had a single evolutionary origin in the Eumaniraptora, or bird-like dinosaurs². This is clearly an exciting finding, and the potential detection of these pigments could indeed enable considerable advances in our understanding of the evolution of egg pigmentation. However, the presence of these two pigments in white avian eggs and crocodile (non-eumaniraptoran) eggs draws their conclusions into question.

The presence of biliverdin and/or protoporphyrin IX does not inevitably mean that an egg is coloured. Both pigments are widespread, and biliverdin can be found in (among many other places) the blood of fish³ and the yolk of frog eggs⁴. When found in eggshells, they can cause brown or blue–green coloration, but many white avian eggs (for example, of chickens (*Gallus gallus*)⁵, white storks (*Ciconia ciconia*), wood pigeons (*Columba palumbus*)⁶ and swifts (*Apus apus*)⁷) contain the pigments in low levels. Indeed, an extensive survey concluded that pigment-free white eggs are rare⁶. An egg containing these pigments could thus either be coloured or not, depending on concentration. Raman data on organic signals, as implemented², can only be interpreted in relative terms, and thus give no information on absolute concentration that in turn would inform on coloration. Furthermore, none of the fossil eggs (with the possible exception of the previously reported *Heyuannia*⁸) shows any evidence of preserved colour. A recent study⁹ suggested that coloured eggs become brown during diagenesis, but did not examine the fate of white eggs. Without such data, it is not possible to infer with certainty that a fossil egg was coloured. Nevertheless, all eggs with any potential signal of preserved pigment were classified as fully coloured and/or colour patterned².

Apparent patterning of these pigments into maculation and spotting, as well as distribution in the shell cross-section reminiscent of those in modern bird eggs could be argued to further support the hypothesis that fossil eggs were coloured. However, these analyses were only performed on fossil eggs already classified as pigmented, making this argument circular. Fossil eggs classified as unpigmented, or white eggs with small amounts of pigment, lacking maculation patterns under the Raman surface-mapping analyses would obviously support the hypothesis that these eggs were coloured. But these critical negative controls were not done. The authors' argument for this omission is that Raman imaging data would only show patterns of background noise. But the same may be true for eggs classified as pigmented. This

possibility is particularly likely given that (1) the wavelengths chosen for Raman imaging analysis ($1,160\text{ cm}^{-1}$ and $1,350\text{ cm}^{-1}$) are outside of the diagnostic Raman 'fingerprint' region, and (2) all 'unpigmented' fossil eggs show strong peaks at $1,160\text{ cm}^{-1}$, whereas most 'pigmented' eggs have valleys, rather than peaks, at $1,350\text{ cm}^{-1}$ (see Extended Data Fig. 1). Thus, these imaging data, although intriguing, provide no further support for colouring of fossil eggs. It could be argued that the low concentrations of pigments found in white eggs are unlikely to be preserved, but this needs to be tested directly and was not done in a related recent study⁹.

Even if we do not definitively know that dinosaur eggs were coloured, it is still interesting and relevant that they may have been physiologically capable of depositing pigments in their eggs. However, the authors' titular conclusion, that this capability had a single origin at the base of eumaniraptorans, is challenged both by the above and even more directly by the recent detection of protoporphyrin in white Siamese crocodile eggs¹⁰ (Fig. 1). Crocodiles are phylogenetically distant from eumaniraptorans, and a new maximum likelihood ancestral state reconstruction that includes them indicates a 67% probability of egg

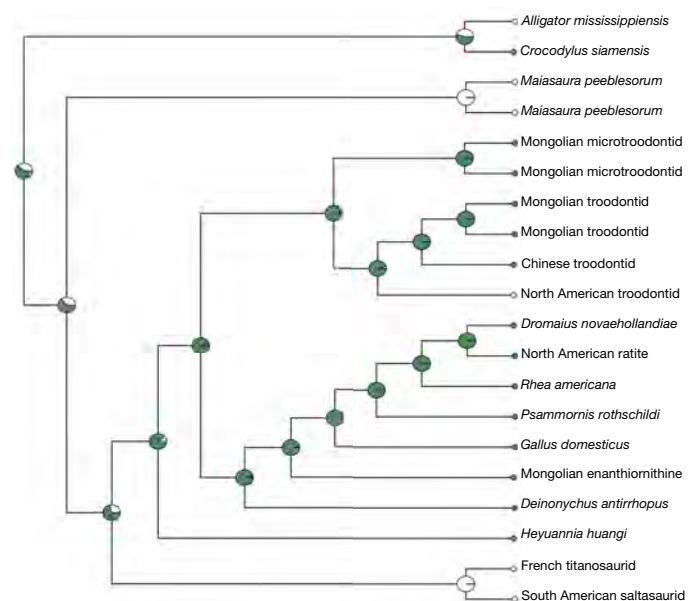


Fig. 1 | Maximum likelihood ancestral reconstruction tree showing probabilities of pigment deposition (green sections of nodes) in eggs of archosaurs. Species and pigment presence data are the same as Wiemann et al.², with the exception of *Crocodylus siamensis*, which is taken from Mikšik et al.¹⁰.

¹Biology Department, Evolution and Optics of Nanostructures Group, University of Ghent, Ghent, Belgium. *e-mail: matthew.shawkey@ugent.be

pigmentation at the base of archosaurs (Fig. 1). Whether extinct crocodilian eggs were coloured or not awaits more quantitative and thorough sampling, but this finding has important implications for the evolution of egg pigmentation. Perhaps deposition of protoporphyrin arose concurrently with hard-shelled eggs, to strengthen shells or serve some other crucial function. Colour (and perhaps biliverdin deposition) may have then arisen later.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

All data were taken from Fig. 1 of Wiemann et al.² with the exception of the data point on Siamese crocodiles, which was based on data from Mikšík et al.¹⁰.

Received: 5 December 2018; Accepted: 17 April 2019;

Published online 19 June 2019.

1. Vinther, J. A guide to the field of palaeo colour. *BioEssays* **37**, 643–656 (2015).
2. Wiemann, J., Yang, T.-R. & Norell, M. A. Dinosaur egg color had a single evolutionary origin. *Nature* **563**, 555–558 (2018).
3. Fang, L.-S. & Bada, J. L. The blue-green blood plasma of marine fish. *Comp. Biochem. Physiol. B* **97**, 37–45 (1990).
4. Marinetti, G. V. & Bagnara, J. T. Yolk pigments of the Mexican leaf frog. *Science* **219**, 985–987 (1983).

5. Baird, T., Solomon, S. E. & Tedstone, D. R. Localisation and characterization of egg shell porphyrin in several avian egg species. *Br. Poult. Sci.* **16**, 201–208 (1975).
6. Kennedy, G. Y. & Vevers, H. G. A survey of avian eggshell pigments. *Comp. Biochem. Physiol. B* **55**, 117–123 (1976).
7. Mikšík, I., Holan, V. & Deyl, Z. Quantification and variability of eggshell pigment content. *Comp. Biochem. Physiol. A* **109**, 769–772 (1994).
8. Wiemann, J. et al. Dinosaur origin of egg color: oviraptors laid blue-green eggs. *PeerJ* **5**, e3706 (2017).
9. Wiemann, J. et al. Fossilization transforms vertebrate hard tissue proteins into N-heterocyclic polymers. *Nat. Comm.* **9**, 4741 (2018).
10. Mikšík, I., Paradis, S., Eckhardt, A. & Sedmera, D. Analysis of Siamese crocodile (*Crocodylus siamensis*) eggshell proteome. *Protein J.* **37**, 21–37 (2018).

Acknowledgements We acknowledge support from FWO grant G007117N and AFOSR grant FA9550-1-18-0447.

Author contributions M.D.S. and L.D. conceived the study, L.D. analysed the data, M.D.S. wrote the manuscript, and both authors edited the manuscript.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-019-1282-4>.

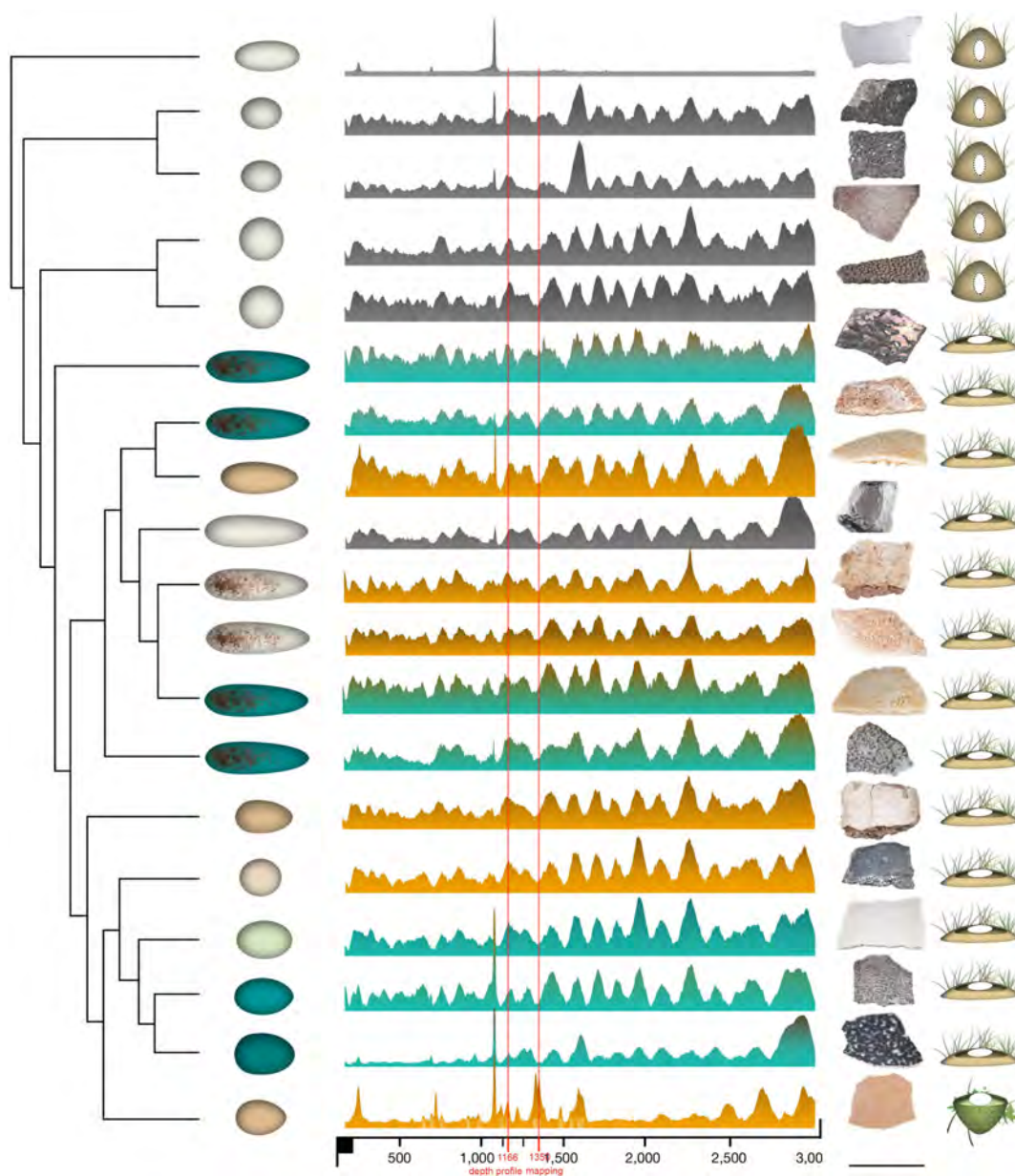
Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-019-1282-4>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to M.D.S.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019



Extended Data Fig. 1 | Figure 1 from Wiemann et al.², with red lines indicating location of wavenumbers analysed using Raman imaging. Both white and coloured eggs have peaks at $1,166\text{ cm}^{-1}$, and coloured eggs have valleys at $1,350\text{ cm}^{-1}$.

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- ☒ ☐ The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- ☒ ☐ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- ☐ ☒ The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- ☒ ☐ A description of all covariates tested
- ☒ ☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- ☐ ☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- ☒ ☐ For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
Give P values as exact values whenever suitable.
- ☒ ☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- ☒ ☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- ☒ ☐ Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection

Data were collected using RStudio Version 1.1.453

Data analysis

All data were analyzed using the packages "ape" (Paradis et al. 2004) and "phytools" (Revell 2012) implemented in the software RStudio Version 1.1.453 (Team, R., 2015. RStudio: integrated development for R. RStudio, Inc., Boston, MA URL <http://www.rstudio.com>).
- Paradis, E., Claude, J. and Strimmer, K., 2004. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics*, 20(2) pp.289-290.
- Revell, L.J., 2012. phytools: an R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution*, 3(2), pp.217-223.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All data are taken from Fig. 1 of the criticized paper (Wiemann et al. (2018)) with the exception of one point (*Crocodylus siamensis*), which was scored as positive based on data in Mikšik et al. (2018).

References:

Mikšik, I., Paradis, S., Eckhardt, A. & Sedmera, D. Analysis of Siamese Crocodile (*Crocodylus siamensis*) Eggshell Proteome. *Protein J.* 37, 21-37 (2018).

Wiemann, J., Yang, T. -R. & Norell, M.A. Dinosaur egg color had a single evolutionary origin. Nature DOI:10.1038/s41586-018-0646-5 (2018).

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☐ Behavioural & social sciences ☒ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	Our comment includes an ancestral reconstruction of egg pigmentation based on the data in Fig. 1 of Wiemann et al. (2018) and one additional data point taken from Mikšík et al. 2018.
Research sample	All data are taken from Fig. 1 of the criticized paper (Wiemann et al. (2018)) with the exception of one point (Crocodylus siamenensis), which was scored as positive based on data in Mikšík et al. (2018).
Sampling strategy	n/a
Data collection	All data are taken from Fig. 1 of the criticized paper (Wiemann et al. (2018)) with the exception of one point (Crocodylus siamenensis), which was scored as positive based on data in Mikšík et al. (2018). All data scored by D'Alba.
Timing and spatial scale	n/a
Data exclusions	No data were excluded from analysis
Reproducibility	n/a
Randomization	n/a
Blinding	n/a
Did the study involve field work?	<input type="checkbox"/> Yes <input checked="" type="checkbox"/> No

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input type="checkbox"/>	<input checked="" type="checkbox"/> Palaeontology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data

Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

Palaeontology

Specimen provenance	n/a
Specimen deposition	n/a
Dating methods	n/a
<input type="checkbox"/> Tick this box to confirm that the raw and calibrated dates are available in the paper or in Supplementary Information.	

Reply to: Egg pigmentation probably has an Archosaurian origin

Jasmina Wiemann^{1*}, Tzu-Ruei Yang^{2,3,5} & Mark A. Norell⁴

REPLYING TO M. D. Shawkey & L. D'Alba *Nature* <https://doi.org/10.1038/s41586-019-1282-4> (2019)

In the accompanying Comment¹, Shawkey and D'Alba suggest that egg pigmentation probably has a single archosaur origin in response to our recent discovery of a single evolutionary origin of dinosaur egg colour². In our study², we analysed the preservation of blue (biliverdin) and red (protoporphyrin) colour pigment in fossilized eggshells of nonavian and avian dinosaurs based on Raman spectroscopy point analyses, protoporphyrin surface maps, and pigment depth profiles. Ornithischian, sauropod and North American troodontid eggs showed no detectable evidence for egg colour pigments, whereas almost all nonavian and avian eumaniraptoran eggs yielded pigment signals². Mapping these results onto an archosaur consensus phylogeny and performing a parsimony-based ancestral state reconstruction revealed a single evolutionary origin of egg colour in (the stem of) eumaniraptoran dinosaurs². Shawkey and D'Alba¹ dispute this conclusion based on several observations: that the presence of eggshell pigments does not necessarily indicate egg colour; that our Raman method cannot quantify pigments and therefore cannot distinguish between coloured and white eggs; that fossil eggs do not show visible evidence of colour; and that we excluded a potentially pigmented yet white *Crocodylus siamensis* eggshell in our ancestral state reconstruction of archosaur egg colour.

Certain eggs are white even if traces of biliverdin and/or protoporphyrin are present³—the intensity of colour reflects the concentration of eggshell pigment⁴. Therefore, as Shawkey and D'Alba¹ point out, dinosaur eggs that yield a pigment signal^{2,5} may have varied from white to intensely coloured (as we discussed and imaged in figure 2a of our Letter²). Raman spectroscopy cannot be used to quantify pigment concentrations^{2,6}, but a crucial amount is necessary to elicit a Raman signal, and the lower spectroscopic detection limit provides information on the concentration present⁶. Exploiting this requirement, we evaluated whether Raman spectroscopy can distinguish between traces of pigment in white eggs and increased amounts in coloured eggs. Additional white and lightly coloured extant eggs (Extended Data Table 1) were analysed under previously published conditions², and the resulting spectra (Fig. 1a) were subjected to two different types of cluster analysis (Fig. 1b). White eggs did not yield a pigment signal (Fig. 1a) because pigment traces did not exceed the detection limit for our Raman-based approach⁶. All lightly coloured eggs, in contrast, produced a pigment signal (Fig. 1a). Both analyses clearly separated a cluster of white eggs from one of lightly coloured eggs (Fig. 1b). We infer that the fossilized dinosaur eggs that yielded a pigment signal² were coloured, as their pigment concentrations exceeded the spectroscopic detection limit⁶.

Shawkey and D'Alba¹ claim that our detection of eggshell spots and speckles using Raman-surface mapping, and our demonstration of pigment depth profiles through vertical sections of eggshell, provide no further support for colour in fossil dinosaur eggs¹. Protoporphyrin causes spots and speckles in eggs^{3,4}, and was mapped

only on fossil eggshell that yielded a protoporphyrin signal²; otherwise, only the nano-differential spectral background noise would have been detected^{2,6}. Both pigments were targeted for the depth profiles². Analysing only samples with a protoporphyrin and/or biliverdin signal present was determined by a technical constraint⁶, rather than a sampling decision that we made². The wavelengths chosen to map out protoporphyrin (1,350 cm⁻¹) and protoporphyrin and biliverdin (1,160 cm⁻¹) were selected to avoid the pigment fingerprint region affected by the signal produced by protein fossilization products (PFPs) and by background fluorescence associated with pigment peaks localized on top of the rather broad PFP spectral shoulder². Figure 1c shows the pigment peak loadings in a chemospace principal component analysis including all the eggshells²; the peak loadings (including background noise) represent their significance for pigment concentration (PC1)² and identification (PC2)². The two chosen peaks are diagnostic of functional groups unique to protoporphyrin (1,350 cm⁻¹)^{2,7} and tetrapyrrol pigments (1,160 cm⁻¹)^{2,7}; they are not affected by a PFP background (Fig. 1c), and represent the most reliable indicator of pigment concentrations and types² (Fig. 1c). Our protoporphyrin maps and pigment depth profiles do not represent noise, but provide evidence of original egg spots and the distribution of pigment across fossil eggshell² (Extended Data Fig. 1). However, it is not possible to reconstruct the original colour of fossil dinosaur eggs in detail, as pigments are lost and/or transformed during diagenesis².

Diagenetic transformation affects not only eggshell pigments^{2,5} but also the eggshell organic matrix⁸, which is originally composed of mucopolysaccharides. These compounds, like all proteinaceous matter in hard tissues exposed to oxidative conditions, form N-heterocyclic polymers, which are responsible for the brown discolouration of eggs during fossilization⁸. This fossilization process affects proteinaceous material in every eggshell (and other vertebrate hard tissue)⁸, regardless of the presence or absence of pigments⁸. The more abundant dark brown protein fossilization products overprint the weak colour generated by minor amounts of unaltered eggshell pigment preserved^{2,8}. The preservation of original colour in fossil dinosaur eggs is practically impossible^{2,8}, and chemical evidence such as that presented in our Letter² is necessary to determine whether a fossil egg was once coloured or not.

Our parsimony-based ancestral state reconstruction² was based on the assumption that detectable pigment reflects originally coloured eggs (as demonstrated in Fig. 1a–c); we coded taxa as either uncoloured (0) or coloured (1). Our analysis revealed a single evolutionary origin of egg colour in eumaniraptoran dinosaurs. Shawkey and D'Alba¹ coded protoporphyrin traces detected by high sensitivity mass spectrometry in a white *Crocodylus siamensis* (crocodile) eggshell⁹ as equivalent to 'coloured' in our fossil eggs as detected by Raman spectroscopy². They added this taxon to our published dataset and ran a maximum likelihood-based ancestral state reconstruction⁹. However, traces of

¹Department of Geology and Geophysics, Yale University, New Haven, CT, USA. ²Institute of Vertebrate Paleontology and Paleoanthropology, Chinese Academy of Sciences, Beijing, China. ³Department of Earth Sciences, National Cheng Kung University, Tainan, Taiwan. ⁴Division of Paleontology, American Museum of Natural History, New York, NY, USA. ⁵Present address: Division of Geology, National Museum of Natural Science, Taichung, Taiwan. *e-mail: jasmina.wiemann@yale.edu

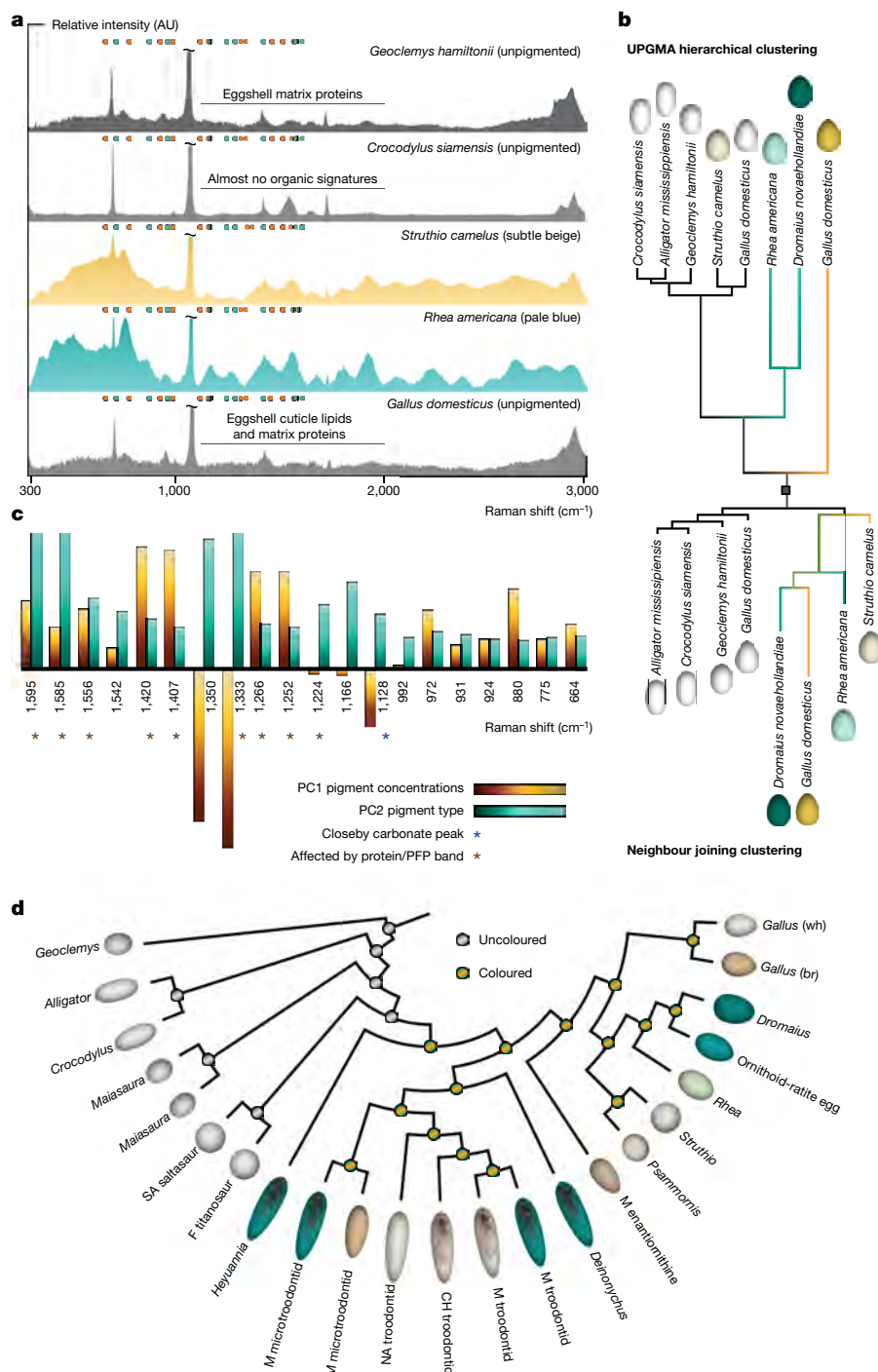


Fig. 1 | Raman spectroscopic and statistical analysis of an eggshell pigmentation versus an egg colour signal. **a**, Raman spectra ($n = 5$) over the wavelength range $300\text{--}3,000\text{ cm}^{-1} \pm 2\text{ cm}^{-1}$ (6 accumulations (technical replicates), 20-s exposure) of *Geoclemys hamiltonii* (turtle), *Crocodylus siamensis* (Siamese crocodile), *Struthio camelus* (ostrich), *Rhea americana* (rhea), and white *Gallus domesticus* (chicken) eggshell (following our previous Letter²). All spectra are baselined and normalized². The colour below the spectral function represents the absence of a pigment signal (grey), the presence of protoporphyrin (orange), or biliverdin (blue). Pigment peak positions are depicted as transparently coloured dots: orange dots indicate a protoporphyrin band, blue dots a biliverdin peak. Increased saturation of dots represents the presence of a pigment peak in the spectrum below. **b**, Cluster analyses for extant coloured ($n = 4$) and uncoloured/white eggs ($n = 4$) based on Raman spectra. Eggshell spectra are clustered by UPGMA hierarchical clustering in the top topology, and by neighbour joining clustering in the bottom topology. The taxa are those we investigated in our Letter², with the additions in **a**. **c**, Loadings plots of all pigment peaks ($n = 20$) based on a principal component (PC) analysis of all eggshells in our Letter² and added in **a**. Orange bars indicate loadings on PC1 separating samples based on pigment concentrations; blue bars indicate loadings on PC2 separating samples based on pigment types. Brown asterisks label pigment peaks that are affected by PFP baseline noise and fluorescence. The blue asterisk labels the $1,128\text{ cm}^{-1} \pm 2\text{ cm}^{-1}$ peak that is affected by the adjacent eggshell carbonate peak. **d**, Parsimony-based ancestral state reconstruction of egg colour (coded as: 0, uncoloured; 1, coloured) across an Archosauromorpha consensus phylogeny² based on the Raman spectroscopic characterization in our Letter², and the additions in **a** ($n = 23$). Grey dots represent uncoloured/white eggs; orange-blue dots represent visibly coloured eggs. Egg icons on the terminal branches represent known egg colour and pattern for extant species, and the reconstructed egg colour and patterns for fossil eggs. A single evolutionary origin of egg colour is found for eumaniraptoran dinosaurs. AU, arbitrary units; br, brown eggshell; CH, Chinese; F, French; M, Mongolian; NA, North American; SA, South American; wh, white eggshell.

protoporphyrin in a white eggshell⁹ do not equate to 'coloured' in our dataset² (and such traces could also derive from haem-containing chorioallantoic vascularity protruding into the basalmost layers of the *Crocodylus siamensis* eggshell, as shown in Extended Data Fig. 1). We analysed eggshells of *Crocodylus siamensis* using Raman spectroscopy under identical conditions to those used for our samples² (Fig. 1a). On this basis, we recovered *Crocodylus siamensis* eggshell as 'uncoloured' (Fig. 1b). We ran a new parsimony-based ancestral state reconstruction based on our original archosaur dataset², together with white eggs of *Geoclemys* (turtle), *Crocodylus siamensis* eggs, and various birds (Fig. 1d); this yielded the same result as before². (Shawkey and D'Alba¹ used maximum likelihood even though the dataset is rather small

owing to biases in the fossil record, and includes non-sequence characters: this method is commonly outperformed by parsimony-based and Bayesian inferences¹⁰.)

The ancestral state reconstruction of Shawkey and D'Alba¹ reveals a single evolutionary origin of 'eggshell pigmentation' (problematic owing to the non-synonymous codings) in archosaurs¹, and they infer that eggshell pigmentation preceded egg colour¹. Our revised parsimony-based approach yields an 'uncoloured' ancestral egg in archosauromorphs, archosaurs, dinosaurs and saurischians (Fig. 1d). Even with the inclusion of various white-shelled avian eggs, we confirm a single evolutionary origin of egg colour in eumaniraptorans (Fig. 1d).

Eggshell pigmentation may represent a more basal trait than true egg colour, but the Raman data on fossil egg colour^{1,2} cannot address this question. The fossil and modern eggshells in our study are reliably coded as coloured or uncoloured. Nonavian eumaniraptoran dinosaur eggs came in various colours and patterns². Incorporating these data into a phylogenetic analysis requires standardized datasets^{2,8}.

Data availability

All data supporting the findings of this study are available within the paper (Fig. 1), and its Extended Data (Extended Data Fig. 1, Extended Data Table 1).

1. Shawkey, M. D. & D'Alba, L. Egg pigmentation probably has an Archosaur origin. *Nature* <https://doi.org/10.1038/s41586-019-1282-4> (2019).
2. Wiemann, J., Yang, T.-R. & Norell, M. A. Dinosaur egg colour had a single evolutionary origin. *Nature* **563**, 555–558 (2018).
3. Kennedy, G. Y. & Vevers, H. G. A survey of avian eggshell pigments. *Comp. Biochem. Physiol. B* **55**, 117–123 (1976).
4. Gosler, A. G., Higham, J. P. & Reynolds, S. J. Why are birds' eggs speckled? *Ecol. Lett.* **8**, 1105–1113 (2005).
5. Wiemann, J. et al. Dinosaur origin of egg color: oviraptors laid blue-green eggs. *PeerJ* **5**, e3706 (2017).
6. Salzer, R. & Siesler, H. W. (eds). *Infrared and Raman Spectroscopic Imaging* (John Wiley & Sons, 2014).
7. Thomas, D. B. et al. Analysing avian eggshell pigments with Raman spectroscopy. *J. Exp. Biol.* **218**, 2670–2674 (2015).

8. Wiemann, J. et al. Fossilization transforms vertebrate hard tissue proteins into N-heterocyclic polymers. *Nat. Commun.* **9**, 4741 (2018).
9. Mikšík, I., Paradis, S., Eckhardt, A. & Sedmera, D. Analysis of Siamese Crocodile (*Crocodylus siamensis*) Eggshell Proteome. *Protein J.* **37**, 21–37 (2018).
10. O'Reilly, J. E. et al. Bayesian methods outperform parsimony but at the expense of precision in the estimation of phylogeny from discrete morphological data. *Biol. Lett.* **12**, 20160081 (2016).

Acknowledgements We thank D. E. G. Briggs for assistance with the manuscript, and M. Fabbri and J. Gauthier for suggestions.

Author contributions J.W., T.-R.Y. and M.A.N. discussed Shawkey and D'Alba's concerns. J.W. designed and performed the experiments, analysed the data, and created the figure. J.W. wrote the manuscript, which was reviewed by all authors.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-019-1283-3>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to J.W.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019



Extended Data Fig. 1 | *Crocodylus siamensis* outer and inner eggshell surfaces. **a**, The outer eggshell surface ($n = 1$), which contains the highest pigment concentrations in eumaniraptorans is uncoloured/white. **b**, **c**, The inner eggshell surface ($n = 2$) shows a reddish tint (**b**), which reflects chorioallantoic tissues and vascularity (**c**) protruding into the innermost layers of the eggshell. The inner eggshell surface in **b** is

manually cleaned of adjacent chorioallantoic tissue, whereas the inner eggshell surface in **c** is left untreated. Chorioallantoic tissues (**c**) are commonly vascularized, and therefore saturated in blood-derived haem, which represents, when dechelated, protoporphyrin. In this case, traces of protoporphyrin recovered from *Crocodylus siamensis* eggshell would not be homologous with eumaniraptoran eggshell protoporphyrin.

MATTERS ARISING

Extended Data Table 1 | Information on added taxa ($n = 4$), catalogue numbers, ages, localities, and egg colours

Assigned Taxon	Catalogue No.	Age	Locality	Colour
<i>Geoclemys hamiltonii</i>	uncatalogued	Extant	Unknown	white
<i>Crocodylus siamensis</i>	YPM HERR 018977	Extant	Myanmar, South East Asia	white
<i>Struthio camelus</i>	YPM ORN 141976	Extant	Egypt, Africa	beige
<i>Gallus domesticus</i>	uncatalogued	Extant	New Haven, CT, USA	white

Antarctic offshore polynyas linked to Southern Hemisphere climate anomalies

Ethan C. Campbell^{1*}, Earle A. Wilson¹, G. W. Kent Moore^{2,3}, Stephen C. Riser¹, Casey E. Brayton⁴, Matthew R. Mazloff⁵ & Lynne D. Talley⁵

Offshore Antarctic polynyas—large openings in the winter sea ice cover—are thought to be maintained by a rapid ventilation of deep-ocean heat through convective mixing. These rare phenomena may alter abyssal properties and circulation, yet their formation mechanisms are not well understood. Here we demonstrate that concurrent upper-ocean preconditioning and meteorological perturbations are responsible for the appearance of polynyas in the Weddell Sea region of the Southern Ocean. Autonomous profiling float observations—collected in 2016 and 2017 during the largest polynyas to form near the Maud Rise seamount since 1976—reveal that the polynyas were initiated and modulated by the passage of severe storms, and that intense heat loss drove deep overturning within them. Wind-driven upwelling of record strength weakened haline stratification in the upper ocean, thus favouring destabilization in 2016 and 2017. We show that previous Weddell polynyas probably developed under similarly anomalous conditions, which are associated with a mode of Southern Hemisphere climate variability that is predicted to strengthen as a result of anthropogenic climate change.

The blanket of sea ice that develops around Antarctica each winter reduces interaction between the ocean and the atmosphere. By eliminating this barrier, large offshore openings in the sea ice pack—sometimes referred to as ‘open-ocean’ or ‘sensible heat’ polynyas—expose the ocean surface to heat extraction by the frigid atmosphere above. The resultant loss of surface buoyancy may drive convective overturning, maintaining ice-free conditions by liberating vast amounts of heat stored precariously just below the cold surface layer^{1–3}. These deep mixing events may rapidly modify the ocean interior, with far-reaching implications for abyssal properties⁴, large-scale ocean circulation^{5–10} and carbon sequestration^{11,12}. Further, heat ventilation during offshore polynya events can be expected to affect the regional atmospheric state^{9,13,14} and possibly global climate patterns through atmospheric teleconnections¹⁵.

The Antarctic Bottom Water that fills the global abyss today originates from the Antarctic continental margin¹⁶. However, in past glacial climates^{16,17}—when grounded ice sheets restricted formation at present-day sites—and perhaps even in pre-industrial times^{18,19}, it may have been formed predominantly in offshore polynyas. This offshore deep water formation pathway is spuriously prevalent in many present-generation climate models, and introduces biases into the present and future Southern Ocean properties, circulation, and sea ice area^{8,20–22}. In contrast to climate models, the modern satellite record (1972–present) shows only intermittent occurrence of offshore polynyas, the largest of which have appeared in the Weddell Sea region near the Maud Rise seamount and in the Cosmonaut Sea offshore of East Antarctica²³. In the most prominent example, early satellite observations revealed massive, persistent Weddell polynyas over three consecutive winters from 1974 to 1976²⁴ (Fig. 1a). Despite an absence of in situ ocean measurements during these events, bottom-reaching mixing within the polynyas was inferred from a hydrographic survey in 1977, which discovered erosion of normal temperature and salinity layering in one location¹.

Shorter-lived, smaller openings have recurred near Maud Rise^{23,25}, with the largest reappearances in 2016 and 2017^{10,26} (Fig. 1). Notably, the 2016–2017 polynyas accompanied a reversal of the positive trend

in Antarctic sea ice extent that has been observed over the past three decades^{27,28}, which raises the possibility that the Maud Rise events may reflect a larger climate signal¹⁰. The tendency of polynyas to emerge near the seamount is not a coincidence. Theory, model simulations and sparse observations suggest that topography–flow interactions at Maud Rise enhance upward heat fluxes and generate eddies that transmit divergent strain to the ice cover^{25,29–32} (see ref. ³² for a comprehensive discussion of related mechanisms). However, the intermittency of openings remains unexplained. More generally, the absence of detailed measurements of an offshore polynya has afforded little opportunity to validate the local processes and climate influences that have been put forward to explain how destabilization, polynya formation and possibly deep overturning may occur within the ice-covered gyres of the Southern Ocean.

Here we present the first comprehensive analysis of the ocean, ice and atmosphere during an offshore Antarctic polynya event. Passive microwave sea ice concentration (SIC) data and ERA-Interim (ERA-I) reanalysis fields depict the evolution of ice and atmosphere in 2016 and 2017 (Figs. 2a, 3a, b), whereas three under-ice profiling floats—purposefully trapped in the rotating Taylor column circulation over Maud Rise³¹ (Fig. 1b, c)—together provide a continuous record from 2011–2018 (Fig. 4). Two of these floats, deployed as part of the Southern Ocean Carbon and Climate Observations and Modeling (SOCCOM) Project, were present during the 2016–2017 polynya events. A synthesis of almost 4,000 past hydrographic observations from floats, ships and instrumented seals near Maud Rise provides year-round climatological baselines, against which 2016 and 2017 may be compared (Extended Data Fig. 1; Methods section ‘Hydrographic climatologies’). Lastly, a new record of past offshore polynyas reveals connections to large-scale climate fluctuations (Fig. 5).

Polynya formation in 2016

On 27 July 2016, a month after ice formed over Maud Rise, a sliver of open water appeared above its northeast flank (Fig. 2a, Extended Data Fig. 2), which is a location that is predisposed for reduced SIC^{25,30,32}.

¹School of Oceanography, University of Washington, Seattle, WA, USA. ²Department of Physics, University of Toronto, Toronto, Ontario, Canada. ³Department of Chemical and Physical Sciences, University of Toronto Mississauga, Mississauga, Ontario, Canada. ⁴School of the Earth, Ocean and Environment, University of South Carolina, Columbia, SC, USA. ⁵Scripps Institution of Oceanography, University of California San Diego, La Jolla, CA, USA. *e-mail: ethancc@uw.edu

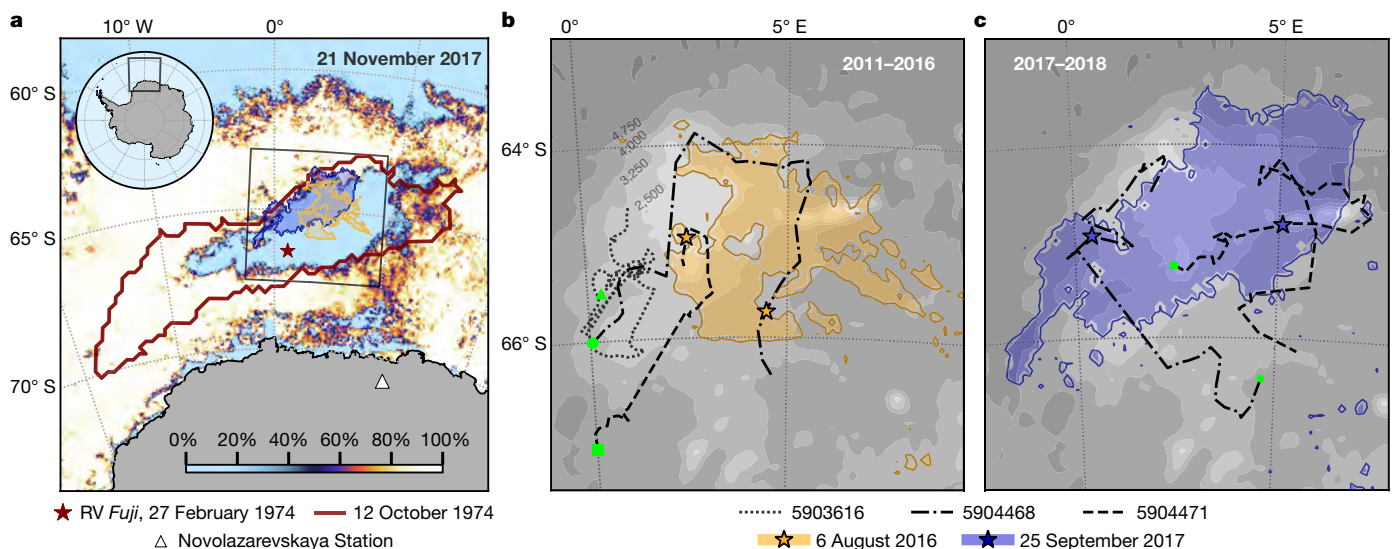


Fig. 1 | Polynyas of 1974, 2016 and 2017 in relation to profiling float trajectories near Maud Rise. a, Sea ice concentration in the eastern Weddell sector of the Southern Ocean on 21 November 2017 from AMSR2-ASI, overlaid with 50% SIC contours showing polynya extent on 12 October 1974 (from Nimbus-5 ESMR; red contour) as well as on 6 August 2016 and 25 September 2017 (from AMSR2-ASI; orange and blue shading). The grey box encompasses the area shown in **b** and **c**. The red star identifies the MLS measurement obtained by research vessel (RV) *Fuji* on 27 February 1974 (shown in Fig. 2b)⁴¹, and the white triangle marks Novolazarevskaya Station, Queen Maud Land (see pressure record in Fig. 5c). **b**, Bathymetry around Maud Rise (depth contours labelled

in metres) with polynya extent from 2016 shaded as in **a**. Trajectories of floats 5903616 (December 2011 to June 2016), 5904468 (January 2015 to December 2016) and 5904471 (December 2014 to December 2016) begin at deployment locations (green) and include estimated locations during the 2016 polynya (orange stars). **c**, Bathymetry as in **b** with polynya extent from 2017 shaded as in **a**. Trajectories of floats 5904468 (January 2017 to May 2018) and 5904471 (January 2017 to June 2018) begin at marked locations (green) and include estimated locations during the 2017 polynya (blue stars). See Extended Data Figs. 2, 8 for full evolution of SIC during the 2016 and 2017 polynyas with float locations.

Despite below-freezing air temperatures, the polynya eventually expanded to 33,000 km² (Fig. 3a, Extended Data Fig. 3b), making it the largest opening from full ice cover since 1976 (Fig. 5a). The polynya closed 21 days later, on 17 August 2016 (Fig. 3a, Extended Data Fig. 2). An exceptionally rapid melt season across all Antarctic sectors followed in November. Although most explanations for this unprecedented sea ice retreat have focused on the variability in atmospheric circulation associated with tropically forced teleconnections (refs^{27,28} and references therein), wind-driven upwelling of warm subsurface water may have also contributed in some regions²⁸.

Profiling float data near Maud Rise show that a late freeze in 2016 (Fig. 2a) followed 4–5 months of increased mixed-layer temperature (Extended Data Fig. 3d) and mixed-layer salinity (MLS; Fig. 2b). These anomalies are consistent with strengthened upwelling and entrainment of Weddell Deep Water, which is warmer and saltier than the overlying upper ocean (Fig. 4a, b). Abnormal ocean conditions persisted into the period of ice formation. Early July featured the two highest MLS measurements in our eastern Weddell record for that month, which correspond to a maximum elevation of MLS above the Maud Rise climatology of +0.12 practical salinity units (psu) (Fig. 2b).

We attribute the 2016 polynya primarily to preconditioning from these salinity anomalies. Under sea ice, stratification is governed by fluctuations in MLS. The special vulnerability of the Maud Rise water column to overturning is in part due to its high climatological MLS relative to that of the surrounding region²⁹ (Fig. 2b). In a normal winter, brine rejection from ice growth over Maud Rise brings the pycnocline to the brink of erosion³³. Complete destabilization, however, is averted through entrainment of warm thermocline water as the mixed layer deepens. Ventilation of this ‘thermal barrier’ into the mixed layer suppresses ice growth, providing a negative feedback that preserves stratification^{3,33–36}. However, by mid-July 2016, an increase in MLS of just 0.05 psu would have eliminated stratification and triggered deep convection (red line in Fig. 2b), less than a third of that required in a normal year.

In this precarious state, we infer that intense storms overrode the stabilizing negative feedback and initiated deep convection by providing two sudden perturbations: ice divergence and enhanced turbulent

mixing. Divergence may occur across a continuum of spatial scales owing to ice deformation and drift³⁷, and enables rapid ice growth and brine rejection while preventing immediate stabilization from ice melt. Wind-driven turbulent mixing entrains heat and salt into the mixed layer³⁴, a response that is amplified under weak stratification³³. The heat may melt ice, as idealized model experiments with storm perturbations have demonstrated for the Maud Rise ice–ocean system³³. Although ice may reform, negating the stabilizing buoyancy input of melt, entrained salt will linger and will reduce stability^{3,34,35}. These perturbations have been observed to co-occur: during a 1994 winter field campaign over Maud Rise, two violent storms elevated mixed-layer temperature by 0.3 °C above freezing and opened leads that forced an evacuation of the researchers’ ice camp³⁸. Young, possibly thinner ice in July 2016 would have been particularly susceptible to wind forcing³⁷, which would enhance both mechanisms.

Using atmospheric reanalysis, we identify the passage of the most severe cyclones near Maud Rise (see Methods section ‘Storm identification’). Most storms coincide with marked reductions in average SIC (Fig. 2a, Extended Data Fig. 4). Indeed, the 2016 polynya opened during a storm that featured 10-m wind speeds of up to 25 m s^{−1} and surface pressures as low as 947 hPa (Fig. 3a, Extended Data Fig. 4a). We find a surprising correspondence ($r = 0.81$) between the evolution of the extent of the polynya in 2016 and the cumulative wind-speed anomaly from a baseline value (Fig. 3a), which indicates that the polynya grew with strong winds from storms and shrank during quiescent periods. Along its ice-covered perimeter, high winds may have triggered ice loss and destabilization through turbulent mixing. Within its interior, heat extraction and salty turbulent entrainment from high winds would have driven convection, preventing ice from reforming. During a storm on 2 August, the polynya expanded concurrently with ocean–atmosphere turbulent heat fluxes of up to 718 W m^{−2} (Fig. 3b, c) and possible ice divergence (Extended Data Fig. 5). A lapse in storm activity and a reduction in heat flux apparently enabled the polynya to close (Fig. 3a, c). Nonetheless, MLS remained near the overturning limit throughout the winter (Fig. 2b). A storm on 28–30 August caused substantial SIC reduction and warm entrainment (Extended Data Figs. 3d, 4a), and in

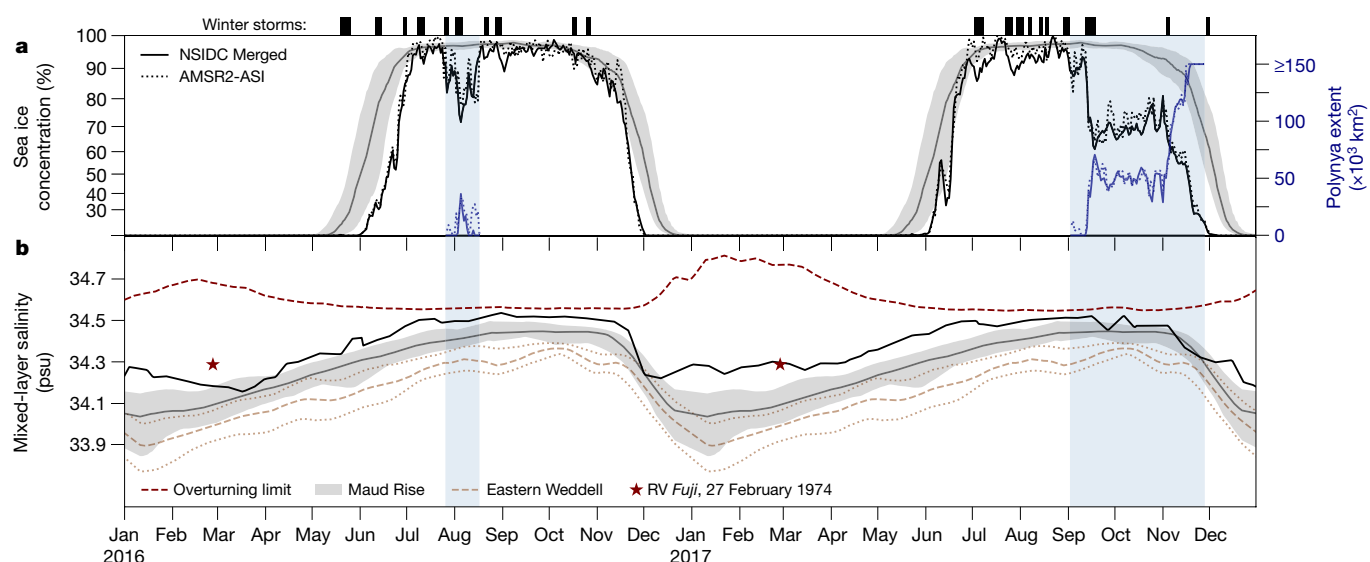


Fig. 2 | Storms, sea ice concentration and mixed-layer salinity at Maud Rise in 2016 and 2017. Marked at the top are intense winter storm events near Maud Rise (also see Extended Data Fig. 4 and Methods section ‘Storm identification’). **a**, Average daily SIC within the Maud Rise region (63° – 67° S, 0° – 10° E) from NSIDC Merged (solid black line) and AMSR2-ASI (dashed black line) in 2016 and 2017. SIC climatology from NSIDC Merged (1978–2019) is shown as median (grey line) and 25%–75% interquartile range (IQR; grey shading). Note the stretched y axis. Polynya extent is quantified (blue lines) during the 2016 and 2017 events (vertical blue shading). **b**, Composite of the highest MLS measured by floats 5903616, 5904468 and 5904471 in 2016 and 2017 (black line;

see Methods sections ‘Derived oceanographic quantities’ and ‘Composites of float time series’). MLS climatology for the Maud Rise region ($R < 250$ km from 65° S, 3° E) is shown as median (grey line) and IQR (grey shading); climatology for the eastern Weddell region away from Maud Rise ($250 < R < 500$ km) is presented for comparison (light brown dashed and dotted lines for median and IQR, respectively; see Methods section ‘Hydrographic climatologies’). Red stars indicate the MLS measured by RV *Fuji* on 27 February 1974 at 66.5° S, 1.2° E, near Maud Rise⁴¹ (see red star in Fig. 1a). The MLS required to eliminate 0–250 m stratification, the ‘overturning limit’, is shown in red (see Methods section ‘Composites of float time series’).

late October—after two storms (Fig. 2a)—a distinct polynya appeared south of Maud Rise, and eventually grew during climatological melt to encompass much of the eastern Weddell (Extended Data Fig. 2).

Deep mixing and resultant preconditioning

We observe an abrupt appearance of anomalies in the ocean interior after the opening of the 2016 polynya: cooling of 0.2°C , freshening of 0.025 psu, and an increase in dissolved oxygen of $10\ \mu\text{mol kg}^{-1}$ at a

depth of around 650 m (Fig. 4a–c, Extended Data Fig. 6). Co-located with these modified patches are bands of near-zero buoyancy frequency (N), which indicate nearly absent stratification (Fig. 4d). These properties reflect recent mixing with winter surface water, as observed during open-ocean convection in the subpolar North Atlantic³⁹. The apparent strengthening and deepening of interior mixed layers after the polynya—observed by both floats (Extended Data Fig. 6)—suggest lateral mixing and homogenization of remnant convective plumes, which probably evolved on scales missed by the coarse sampling in space and time (see Methods section ‘Derived oceanographic quantities’). The isolated intrusion from 500 m to 800 m, which is embedded within displaced isotherms and isohalines (Fig. 4a, b), resembles the submesoscale coherent vortices that are frequently observed after open-ocean convection, such as in the Mediterranean Sea⁴⁰. Together these fingerprints identify the altered water mass as the product of deep-reaching convective mixing.

Analysis of the reduced deep heat content in post-polynya profiles yields estimated ventilation rates (Methods section ‘Polynya heat flux estimates’) that approach the average ocean–atmosphere turbulent heat flux during the 2016 polynya ($252\ \text{W m}^{-2}$; Fig. 3c, Extended Data Fig. 7). This close correspondence suggests that deep convection was driven primarily by a continuous loss of surface buoyancy, rather than by the interior thermobaric mixing that has been invoked to explain overturning near Maud Rise³⁶. Persistent above-freezing mixed-layer temperature provides additional evidence for continuous thermal convection (Extended Data Fig. 3d). Nonetheless, thermobaricity may have modulated the depth of mixing by supplying additional kinetic energy to plumes descending across the pycnocline.

Crucially, the deep freshening that accompanied heat loss in 2016 implies a net upwards mixing of salt⁴¹, thus reducing stability in 2017. This upwards transfer is corroborated by an increase in total salt content from 0 m to 250 m—equivalent to the effect of brine rejection from 0.46 m of ice formation (Extended Data Fig. 3e)—as well as an average MLS elevation of 0.15 psu over the ice-free months of 2017 (Fig. 2b).

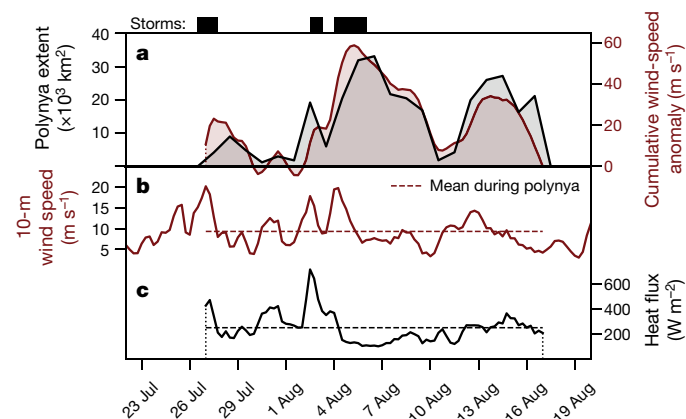


Fig. 3 | Local meteorology and heat loss during the 2016 polynya. Intense winter storm events near Maud Rise are identified at the top, as in Fig. 2. **a**, Daily polynya extent from AMSR2-ASI (see Methods section ‘Polynya identification’) is shaded in black. Overlaid in red is the cumulative sum of the anomaly of 10-m wind speed (as shown in **b**) from its mean value during the 2016 polynya. **b**, Average six-hourly 10-m wind speed from ERA-I reanalysis within the Maud Rise region (63° – 67° S, 0° – 10° E). **c**, Average six-hourly ocean–atmosphere turbulent heat flux within the polynya (see Methods section ‘Atmospheric reanalysis’). Dashed lines indicate mean values during the polynya event.

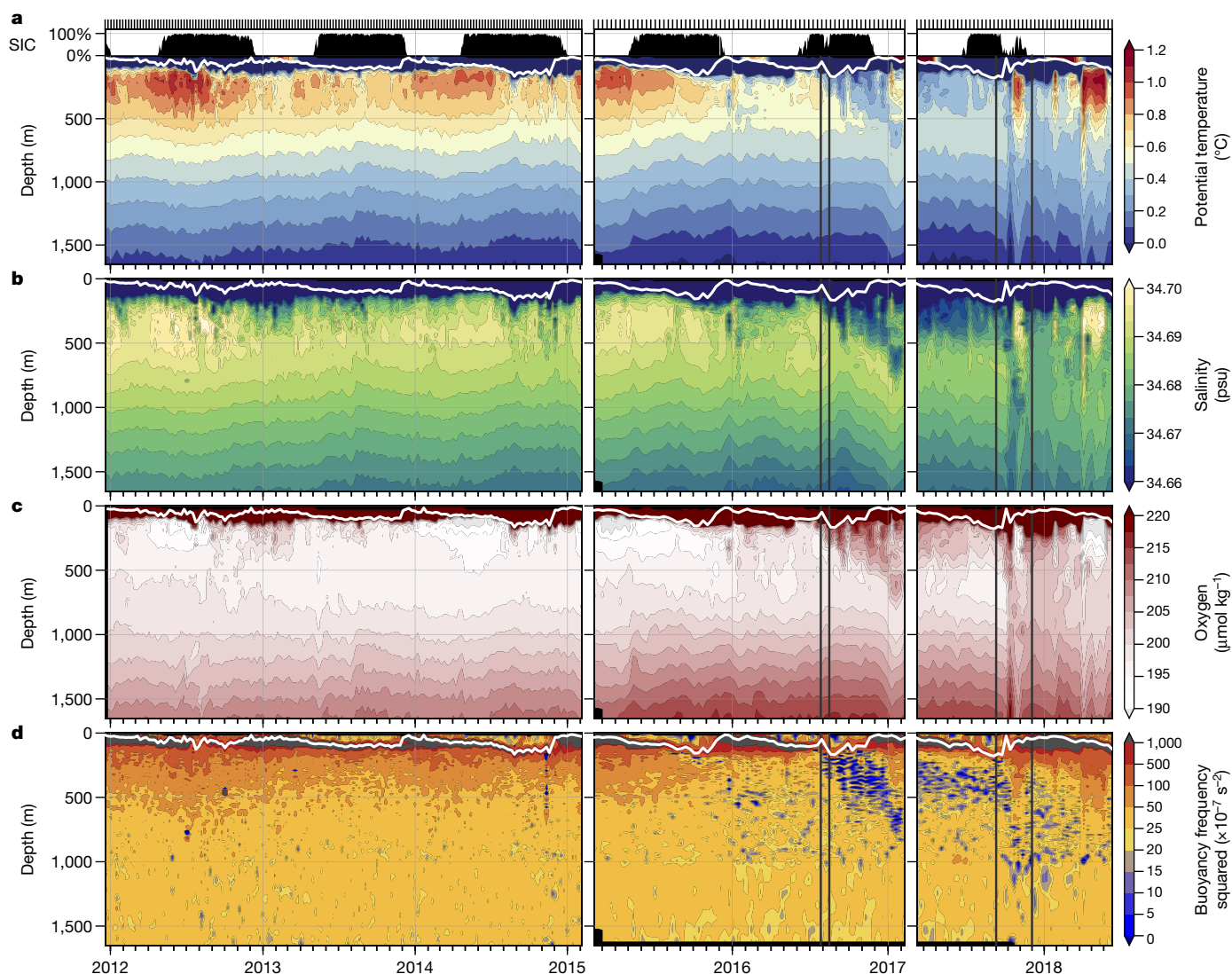


Fig. 4 | Hydrographic observations from Maud Rise from 2011–2018. **a–d**, Continuous-time depth sections of potential temperature (**a**), salinity (**b**), dissolved oxygen (**c**) and buoyancy frequency squared (N^2) (**d**) assembled using observations from profiling floats 5903616 (left), 5904468 (centre) and 5904471 (right). Individual profiles are marked at the top

(black ticks). Mixed-layer depth is indicated in white. Vertical lines in each panel mark the start and end dates of the 2016 and 2017 polynyas. Along-trajectory SIC, primarily from AMSR2-ASI, is shaded at the top in black (see Methods section ‘Sea ice concentration data’). See Extended Data Fig. 6 for the complete time range of data from each float.

This extreme degree of preconditioning is comparable to that observed near Maud Rise before the 1974 polynya⁴¹ (red stars in Figs. 1a, 2b).

Polynya recurrence in 2017

Low upper-ocean stability in early 2017 preconditioned the region for the emergence of a larger polynya than that formed in 2016 (Fig. 1). In July and August, SIC and upper-ocean salt content were highly variable (Fig. 2a, Extended Data Fig. 3e), indicating cycles of melt and refreeze consistent with a strengthened negative feedback to ice growth and a weakly stabilizing buffer from thin ice. Indeed, we estimate that net in situ ice growth in 2017 was half of that experienced in a normal winter (Extended Data Fig. 3e). On 27 July, after four days of storm activity, numerous scattered, short-lived polynyas appeared southwest of Maud Rise. Later, on 3 September, two larger openings emerged above the seamount during a brief storm; they then merged and rapidly grew after storms from 13 to 18 September, which featured the highest wind speeds seen in reanalysis that winter (Fig. 2a, Extended Data Figs. 4b, 8). Within this expanding polynya, both floats inferred ice-free conditions from prolonged warm near-surface temperatures and surfaced to transmit data—a highly unusual occurrence (Extended Data Fig. 8; see Methods section ‘Hydrographic data’). A 1.5-month lull in storm

activity saw the extent of the polynya stabilize at around 50,000 km², before seasonal ice melt and coalescence with the open ocean occurred in November (Fig. 1a, c, 2a, Extended Data Fig. 8).

In October 2017, a near-homogeneous cold, fresh and high-oxygen coherent structure, spanning depths from around 200 m to 1,700 m, was detected from within the polynya. The development of such a structure indicates deeper and more vigorous convection than in 2016 (Fig. 4a–c). As with the remnant ‘chimney’ observed in 1977¹, normal ocean layering was absent. Whether mixing extended below 1,700 m in 2017 cannot be determined from the float data, but we note that the low interior stratification offers little resistance to convection (Extended Data Fig. 6e). This raises the possibility that smaller, shorter-lived polynyas may introduce modification to a similar depth as the 1974–1976 events. However, mixing within transient polynyas near Maud Rise need not be bottom-reaching to affect the properties of the abyssal world ocean. Counterintuitively, bottom water masses remain mostly topographically confined to the Weddell Basin, whereas deep and intermediate layers outflow and spread northwards as Antarctic Bottom Water (refs^{42,43} and references therein). Even brief polynyas may have an outsized effect. Total heat loss during the 21-day opening in 2016 was approximately equal to that which occurs during a

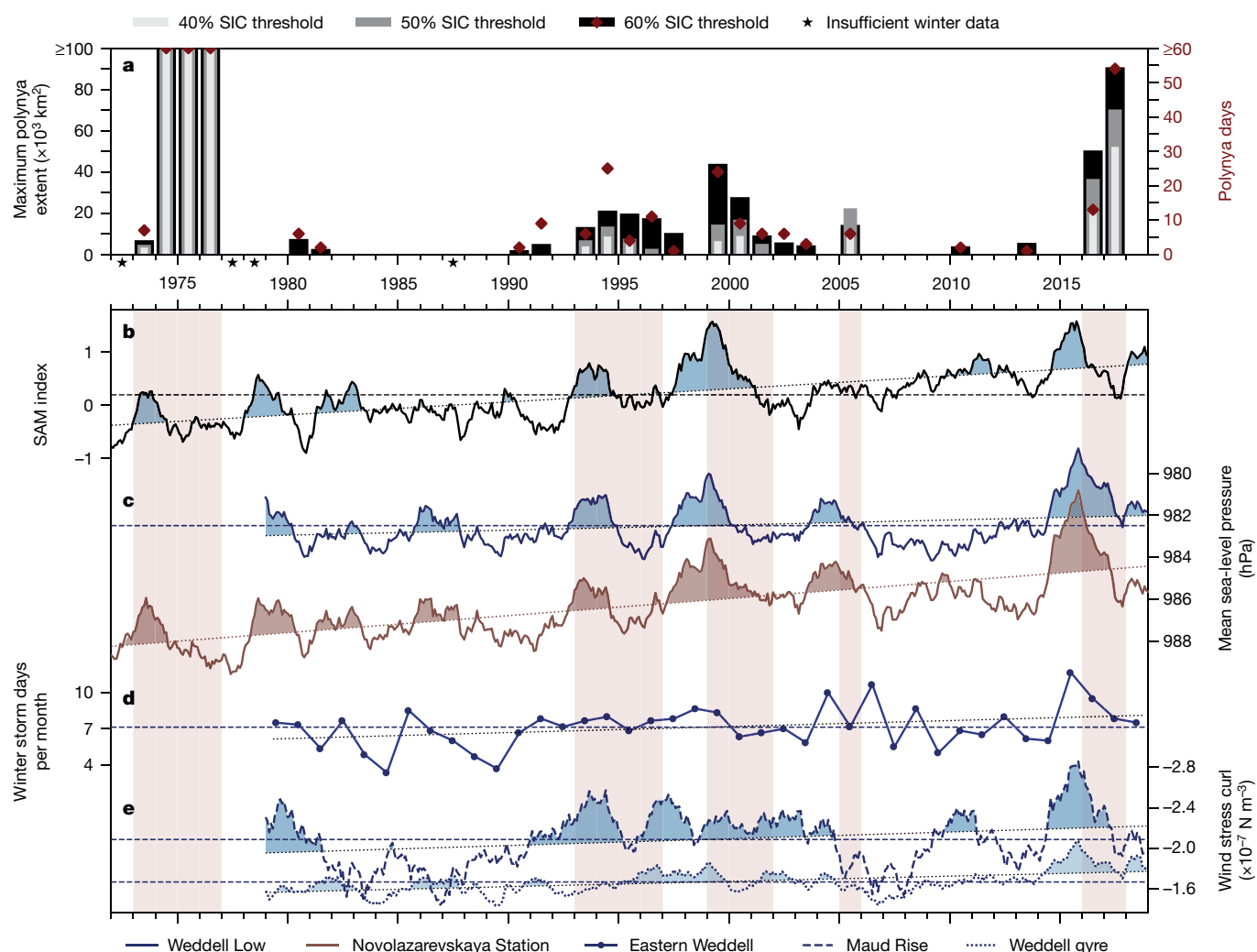


Fig. 5 | Relationships between past polynyas near Maud Rise and climate forcing from 1972–2018. **a**, Annual maximum polynya extent (bars) and number of polynya days (red diamonds; see Methods section ‘Polynya identification’). Maximum polynya extent is calculated for three SIC thresholds representing increasingly strict polynya definitions: 60%, 50% and 40%. Polynya days are quantified using the 60% threshold. Stars indicate years with incomplete or absent SIC records. **b**, SAM index for the years 1972–2018. **c**, Mean sea-level pressure from ERA-I reanalysis, for 1979–2018, within the Weddell Low region (blue; see Methods section ‘Weddell Low’), and from Novolazarevskaya Station, Queen Maud Land, for 1972–2018 (brown; see white triangle in Fig. 1a and Methods section

‘Meteorological station records’). Note the reversed y axis. **d**, Eastern Weddell region (62°–68° S, 15° W–20° E) winter (May–October) mean storm days per month (see Methods section ‘Storm identification’). **e**, Average wind stress curl from ERA-I over Maud Rise (63°–67° S, 0°–10° E; dashed) and the entire Weddell gyre (60° S to the Antarctic continent, 60° W–45° E; dotted). Time series in **b**, **c** and **e** are filtered using a two-year centred running mean and shaded above their linear trends to highlight longer-term fluctuations. See Extended Data Table 1 for trends and significance for **b**–**e**. Horizontal dashed lines are mean values. Years with polynya activity at the 50% threshold are shaded vertically in red in **b**–**e**.

normal ice-covered winter²⁹, thus accelerating the transformation of circumpolar-derived Weddell Deep Water on its slow transit through the Weddell gyre⁴³.

Role of climate variability

Although 2016 and 2017 were exceptional, transient polynyas near Maud Rise have also occurred in clusters of years from 1993–1996 and 1999–2001, as well as in 2005 (Fig. 5a). This is in line with sequences of polynya years found in a high-resolution model simulation of the region³². We have shown that the recurrence of polynyas may be explained by upward salt transfer from convective mixing. However, what controls the intermittency of initial openings is still unknown. By examining climate records from 1972–2018 to identify a common mechanism, we find that the Southern Annular Mode (SAM)—the leading mode of variability in the Southern Hemisphere⁴⁴—fluctuates in lockstep (Fig. 5b, c, Extended Data Table 1) with indices of mean sea-level pressure for the Weddell Low, the climatological low surface pressure centre east of Maud Rise ($r = -0.71$; see Methods section

‘Weddell Low’), and from Novolazarevskaya Station, south of Maud Rise ($r = -0.82$; see white triangle in Fig. 1a). In turn, the curls of wind stress over both the entire Weddell gyre and Maud Rise are correlated with the Weddell Low ($r = 0.61$ and 0.48 , respectively; Fig. 5c, e). Also correlated with the Weddell Low is the frequency of severe winter storms within the eastern Weddell region ($r = -0.55$; Fig. 5d). These relationships are analogous to those associated with variability of the Amundsen Sea Low⁴⁵.

Figure 5 illustrates that past transient polynyas near Maud Rise were preceded by synchronous deviations in these indices: positive SAM, deeper Weddell Low, strengthened cyclonic wind stress curl and frequent winter storm activity. Notably, there were more than twice as many winter storms in 2015 and 2016 as occurred in the least stormy years (Fig. 5d). We have suggested that intense storms facilitate ice loss through turbulent mixing and ice divergence. The other large-scale anomalies provide the preconditioning necessary to permit a polynya in the first place². Local wind stress curl is a proxy for ocean upwelling velocity due to the Ekman relation, and modelling and

theoretical studies indicate that enhanced upwelling favours polynya development by concentrating warm, salty Weddell Deep Water closer to the surface^{2,6,32,33,35,46}. Perhaps more important, however, are the salinity fluxes into the mixed layer that are associated with upwelling. Record-strong wind stress curl in 2015 and 2016 (Fig. 5e) increased MLS near Maud Rise by at least 0.06 psu (Methods section ‘Salinity fluxes from upwelling’), which reduced upper-ocean stability⁴¹ and preconditioned the 2016 opening. Spin-up of the barotropic Weddell gyre circulation due to strengthened wind stress curl (see Methods section ‘Weddell Low’) may have also enhanced topography- or eddy-related upwelling^{29,31} and possibly topographic mixing as faster flow impinged on Maud Rise. Either could have intensified upward heat fluxes in 2016, contributing to the delayed freeze and presumably thinner ice in July. Ultimately, these upwelling- and storm-induced offshore polynyas are linked to positive fluctuations in SAM, which represent large-scale anomalies in the climate of the Southern Hemisphere.

The exceptional 1974–1976 Weddell polynyas, which were larger than the 2016–2017 openings (Fig. 1a), do not conform perfectly to this explanation. The 1974 event occurred under similar circumstances to the 2017 polynya, in that it was preceded by a positive deviation in SAM (Fig. 5b) and lower mean sea-level pressure (Fig. 5c), as well as delayed ice formation in 1973 (not shown) and a brief ‘precursor’ polynya that year^{2,23} (Fig. 5a). But these anomalies were modest. The degree and spatial extent of preconditioning in 1974 were probably augmented by previously proposed mechanisms: a generally saltier Southern Ocean surface¹⁹ and below-average precipitation over the preceding decade associated with prolonged negative SAM⁴⁷. Reduced precipitation, however, cannot explain the 2016 opening, which followed six years of higher-than-average precipitation (Extended Data Fig. 9c).

Comparison with climate models

Accurate modelling of the Southern Ocean requires realistic simulation of deep water formation and ventilation processes²⁰. This remains an urgent challenge. In many climate models that are included in the fifth Coupled Model Intercomparison Project (CMIP5), large Weddell polynyas featuring open-ocean deep convection recur on widely varying timescales^{8,19}. In these models and others, heat steadily accumulates in Weddell Deep Water over several years or decades^{5,10,22,48}, possibly owing to freshwater forcing biases and insufficient vertical mixing^{7,21}. This build-up eventually erodes stratification in an episode of spontaneous ventilation; cycles of recharge and discharge of heat then follow, associated with Weddell polynyas^{5,8,10,22,48}. Observational records of ocean and sea ice are insufficiently long to assess whether these convective cycles, which create a global mode of internal variability^{5,8–10,12,15}, are natural or spurious. This is concerning, because some models suggest that recent Southern Ocean sea surface temperature and sea ice trends may simply reflect this cycle—that is, represent a rebound from ventilation during the 1974–1976 Weddell polynyas¹⁰.

The 2016–2017 events offer an opportunity to assess whether destratification from mid-depth heat accumulation, which is prevalent in models, contributes to polynya formation in reality. The preceding decade-long hiatus in polynya activity is presumably long enough to discern trends in subsurface heat (Fig. 5a). However, we do not find marked heat accumulation before 2016 in records compiled from about 3,000 sub-pycnocline temperature measurements within the eastern Weddell region (Extended Data Fig. 9d, e, Extended Data Table 1; Methods section ‘Sub-pycnocline temperature records’). Although the presence of warm Weddell Deep Water is required to sustain overturning, an increase in local heat content is probably not the immediate cause of polynyas near Maud Rise. This suggests that model results involving Weddell polynyas should be interpreted with greater caution, even when the simulated features resemble observed polynyas in size and location.

Similarly, although deep convection in simulated Weddell polynyas often proceeds until the subsurface heat reservoir is depleted^{5,8,10,48}, the observed cooling after the 1974–1976 and 2016–2017 polynyas of up to 1.0 °C (ref. 1; see Methods section ‘Sub-pycnocline temperature records’) and 0.3 °C (Fig. 4a), respectively, is probably not severe enough to

explain the lack of major openings in 1977²⁴ and 2018 (Fig. 5a). Instead, reductions in upwelling (Fig. 5e) and storm frequency (Fig. 5d), combined with westward advection of upper-ocean salinity anomalies away from Maud Rise^{2,32}, probably terminate multiyear polynya sequences.

Future implications

Progress in simulating and predicting offshore Antarctic polynyas will require a more precise understanding of how storms cause sea ice destruction and how upwelling affects haline stratification. New insight into these relationships may also inform long-standing questions on how Southern Ocean stratification modulates deep ventilation on glacial–interglacial timescales⁴⁹; here, the possible role of offshore polynyas is poorly understood^{16,17}. Although observations of additional offshore polynyas would be valuable in clarifying these mechanisms, climate models predict that their formation may soon be prohibited by the Southern Ocean surface freshening trend of about 0.01 psu per decade, which is associated with anthropogenic climate change¹⁹. However, this simulated non-convective regime has evidently not yet been reached. At present, we observe that the Maud Rise region experiences interannual variability in upper-ocean salinity—and thus also stability—that is an order of magnitude higher than this slow decadal freshening rate. This may enable intermittent openings similar to those in 2016 and 2017 for decades to come. Moreover, we identify statistically significant positive multidecadal trends in polynya-favourable conditions: the strengthening of SAM, Weddell Low, cyclonic wind stress curl and winter storm activity (Extended Data Table 1). These reflect poleward shifts in Southern Hemisphere westerly winds and storm tracks, which are expected to continue under anthropogenic forcing^{44,50}. We propose that these changes may bring enhanced upwelling and a more frequently disturbed sea ice cover to the Weddell Sea region, conceivably signalling a greater future role for transient offshore polynyas in opening a window to the abyssal ocean.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-019-1294-0>.

Received: 14 January 2019; Accepted: 2 May 2019;

Published online 10 June 2019.

- Gordon, A. L. Deep Antarctic convection west of Maud Rise. *J. Phys. Oceanogr.* **8**, 600–612 (1978).
- Martinson, D. G., Killworth, P. D. & Gordon, A. L. A convective model for the Weddell Polynya. *J. Phys. Oceanogr.* **11**, 466–488 (1981).
- Martinson, D. G. Evolution of the Southern Ocean winter mixed layer and sea ice: Open ocean deepwater formation and ventilation. *J. Geophys. Res.* **95**, 11641–11654 (1990).
- Zanowski, H., Hallberg, R. & Sarmiento, J. L. Abyssal ocean warming and salinification after Weddell polynyas in the GFDL CM2G coupled climate model. *J. Phys. Oceanogr.* **45**, 2755–2772 (2015).
- Martin, T., Park, W. & Latif, M. Multi-centennial variability controlled by Southern Ocean convection in the Kiel Climate Model. *Clim. Dyn.* **40**, 2005–2022 (2013).
- Cheon, W. G., Park, Y.-G., Toggweiler, J. R. & Lee, S.-K. The relationship of Weddell Polynya and open-ocean deep convection to the Southern Hemisphere westerlies. *J. Phys. Oceanogr.* **44**, 694–713 (2014).
- Heuzé, C., Ridley, J. K., Calvert, D., Stevens, D. P. & Heywood, K. J. Increasing vertical mixing to reduce Southern Ocean deep convection in NEMO3.4. *Geosci. Model Dev.* **8**, 3119–3130 (2015).
- Behrens, E. et al. Southern Ocean deep convection in global climate models: A driver for variability of subpolar gyres and Drake Passage transport on decadal timescales. *J. Geophys. Res. Oceans* **121**, 3905–3925 (2016).
- Pedro, J. B. et al. Southern Ocean deep convection as a driver of Antarctic warming events. *Geophys. Res. Lett.* **43**, 2192–2199 (2016).
- Zhang, L., Delworth, T. L., Cooke, W. & Yang, X. Natural variability of Southern Ocean convection as a driver of observed climate trends. *Nat. Clim. Change* **9**, 59–65 (2019).
- Bernardello, R., Marinov, I., Palter, J. B., Galbraith, E. D. & Sarmiento, J. L. Impact of Weddell Sea deep convection on natural and anthropogenic carbon in a climate model. *Geophys. Res. Lett.* **41**, 7262–7269 (2014).

12. Resplandy, L., Séférian, R. & Bopp, L. Natural variability of CO₂ and O₂ fluxes: what can we learn from centuries-long climate models simulations? *J. Geophys. Res. Oceans* **120**, 384–404 (2015).
13. Moore, G. W. K., Alverson, K. & Renfrew, I. A. A reconstruction of the air–sea interaction associated with the Weddell polynya. *J. Phys. Oceanogr.* **32**, 1685–1698 (2002).
14. Weijer, W. et al. Local atmospheric response to an open-ocean polynya in a high-resolution climate model. *J. Clim.* **30**, 1629–1641 (2017).
15. Cabré, A., Marinov, I. & Gnanadesikan, A. Global atmospheric teleconnections and multidecadal climate oscillations driven by Southern Ocean convection. *J. Clim.* **30**, 8107–8126 (2017).
16. Amblas, D. & Dowdeswell, J. A. Physiographic influences on dense shelf-water cascading down the Antarctic continental slope. *Earth Sci. Rev.* **185**, 887–900 (2018).
17. Smith, J. A., Hillenbrand, C.-D., Pudsey, C. J., Allen, C. S. & Graham, A. G. C. The presence of polynyas in the Weddell Sea during the Last Glacial Period with implications for the reconstruction of sea-ice limits and ice sheet history. *Earth Planet. Sci. Lett.* **296**, 287–298 (2010).
18. Broecker, W. S., Sutherland, S. & Peng, T.-H. A possible 20th-century slowdown of Southern Ocean deep water formation. *Science* **286**, 1132–1135 (1999).
19. de Lavergne, C., Palter, J. B., Galbraith, E. D., Bernardello, R. & Marinov, I. Cessation of deep convection in the open Southern Ocean under anthropogenic climate change. *Nat. Clim. Change* **4**, 278–282 (2014).
20. Heuzé, C., Heywood, K. J., Stevens, D. P. & Ridley, J. K. Southern Ocean bottom water characteristics in CMIP5 models. *Geophys. Res. Lett.* **40**, 1409–1414 (2013).
21. Kjellsson, J. et al. Model sensitivity of the Weddell and Ross seas, Antarctica, to vertical mixing and freshwater forcing. *Ocean Model.* **94**, 141–152 (2015).
22. Reintges, A., Martin, T., Latif, M. & Park, W. Physical controls of Southern Ocean deep-convection variability in CMIP5 models and the Kiel Climate Model. *Geophys. Res. Lett.* **44**, 6951–6958 (2017).
23. Comiso, J. C. & Gordon, A. L. Recurring polynyas over the Cosmonaut Sea and the Maud Rise. *J. Geophys. Res.* **92**, 2819–2833 (1987).
24. Carsey, F. D. Microwave observation of the Weddell polynya. *Mon. Weath. Rev.* **108**, 2032–2044 (1980).
25. Lindsay, R. W., Holland, D. M. & Woodgate, R. A. Halo of low ice concentration observed over the Maud Rise seamount. *Geophys. Res. Lett.* **31**, L13302 (2004).
26. Swart, S. et al. Return of the Maud Rise polynya: climate litmus or sea ice anomaly? [in “State of the Climate in 2017”]. *Bull. Am. Meteorol. Soc.* **99**, S188–S189 (2018).
27. Wang, G. et al. Compounding tropical and stratospheric forcing of the record low Antarctic sea-ice in 2016. *Nat. Commun.* **10**, 13 (2019).
28. Meehl, G. A. et al. Sustained ocean changes contributed to sudden Antarctic sea ice retreat in late 2016. *Nat. Commun.* **10**, 14 (2019).
29. Gordon, A. L. & Huber, B. A. Southern Ocean winter mixed layer. *J. Geophys. Res.* **95**, 11655–11672 (1990).
30. Holland, D. M. Explaining the Weddell Polynya—a large ocean eddy shed at Maud Rise. *Science* **292**, 1697–1700 (2001).
31. de Steur, L., Holland, D. M., Muench, R. D. & McPhee, M. G. The warm-water “Halo” around Maud Rise: properties, dynamics and impact. *Deep Sea Res. Part I* **54**, 871–896 (2007).
32. Kurtakoti, P., Veneziani, M., Stössel, A. & Weijer, W. Preconditioning and formation of Maud Rise polynyas in a high-resolution earth system model. *J. Clim.* **31**, 9659–9678 (2018).
33. Wilson, E. A., Riser, S. C., Campbell, E. C. & Wong, A. P. S. Winter upper-ocean stability and ice–ocean feedbacks in the sea ice-covered Southern Ocean. *J. Phys. Oceanogr.* **49**, 1099–1117 (2019).
34. Martinson, D. G. & Iannuzzi, R. A. in *Antarctic Sea Ice: Physical Processes, Interactions and Variability (Antarctic Research Series)* Vol. 74 (ed. Jeffries, M. O.) 243–271 (American Geophysical Union, 1998).
35. Timmermann, R., Lemke, P. & Kottmeier, C. Formation and maintenance of a polynya in the Weddell Sea. *J. Phys. Oceanogr.* **29**, 1251–1264 (1999).
36. McPhee, M. G. Marginal thermobaric stability in the ice-covered upper ocean over Maud Rise. *J. Phys. Oceanogr.* **30**, 2710–2722 (2000).
37. Itkin, P. et al. Thin ice and storms: sea ice deformation from buoy arrays deployed during N-ICE2015. *J. Geophys. Res. Oceans* **122**, 4661–4674 (2017).
38. McPhee, M. G. et al. The Antarctic Zone Flux Experiment. *Bull. Am. Meteorol. Soc.* **77**, 1221–1232 (1996).
39. Våge, K. et al. Surprising return of deep convection to the subpolar North Atlantic Ocean in winter 2007–2008. *Nat. Geosci.* **2**, 67–72 (2009).
40. Testor, P. et al. Multiscale observations of deep convection in the northwestern Mediterranean Sea during winter 2012–2013 using multiple platforms. *J. Geophys. Res. Oceans* **123**, 1745–1776 (2018).
41. Motoi, T., Ono, N. & Wakatsuchi, M. A mechanism for the formation of the Weddell Polynya in 1974. *J. Phys. Oceanogr.* **17**, 2241–2247 (1987).
42. Mantyla, A. W. & Reid, J. L. Abyssal characteristics of the World Ocean waters. *Deep-Sea Res. A* **30**, 805–833 (1983).
43. Jullion, L. et al. The contribution of the Weddell Gyre to the lower limb of the Global Overturning Circulation. *J. Geophys. Res. Oceans* **119**, 3357–3377 (2014).
44. Thompson, D. W. J. et al. Signatures of the Antarctic ozone hole in Southern Hemisphere surface climate change. *Nat. Geosci.* **4**, 741–749 (2011).
45. Fogt, R. L., Wovrosh, A. J., Langen, R. A. & Simmonds, I. The characteristic variability and connection to the underlying synoptic activity of the Amundsen-Bellinghousen Seas Low. *J. Geophys. Res. Atmos.* **117**, D07111 (2012).
46. Cheon, W. G. et al. Replicating the 1970s’ Weddell Polynya using a coupled ocean–sea ice model with reanalysis surface flux fields. *Geophys. Res. Lett.* **42**, 5411–5418 (2015).
47. Gordon, A. L., Visbeck, M. & Comiso, J. C. A possible link between the Weddell Polynya and the Southern Annular Mode. *J. Clim.* **20**, 2558–2571 (2007).
48. Dufour, C. O. et al. Preconditioning of the Weddell Sea polynya by the ocean mesoscale and dense water overflows. *J. Clim.* **30**, 7719–7737 (2017).
49. Sigman, D. M., Hain, M. P. & Haug, G. H. The polar ocean and glacial cycles in atmospheric CO₂ concentration. *Nature* **466**, 47–55 (2010).
50. Chang, E. K. M., Guo, Y. & Xia, X. CMIP5 multimodel ensemble projection of storm track change under global warming. *J. Geophys. Res. Atmos.* **117**, D23118 (2012).

Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

METHODS

Regions. Various gridded fields were averaged within the Maud Rise region (63°–67° S, 0°–10° E), which loosely encompasses the 4,000 m isobath around the seamount (Fig. 1b, c) and the area of polynya formation in 2016 and 2017. Other metrics were calculated within an area that we designate the eastern Weddell region (62°–68° S, 15° W–20° E), which spans the eastern Weddell, Lazarev and western Riiser-Larsen seas. This larger area increases signal strength for records with spatial uncertainty or sparsity, such as discrete polynya events, storm statistics and precipitation.

In the context of hydrographic observations, the Maud Rise region instead refers to a radius of 250 km from Maud Rise (65° S, 3° E; see Extended Data Fig. 1). This encloses the ‘halo’ of elevated subsurface heat content identified over the seamount^{31,51}. Here the eastern Weddell region refers to hydrographic observations collected within 500 km of Maud Rise.

Sea ice concentration data. Satellite observations of Antarctic sea ice before 1972 are presently limited to a visible-band composite of September 1964 from the NASA Nimbus I mission, which showed a possible offshore polynya in the Weddell Sea⁵². The quality of the imagery prevents conclusive identification. The modern passive microwave SIC era began in 1972 with the single-channel Nimbus-5 Electrically Scanning Microwave Radiometer (ESMR)²⁴. We use daily SIC from December 1972 to May 1977 from the NSIDC Nimbus-5 ESMR v1 product on a 25-km polar stereographic grid, the result of heavy reprocessing⁵³. Gaps of weeks to months are present.

A series of more reliable multi-channel sensors followed. We use the merged NASA Goddard Space Flight Center v3 25-km product distributed by NSIDC from November 1978 to December 2017⁵⁴. The merged Goddard product is nearly identical to the NOAA/NSIDC Climate Data Record (CDR), which is based on two well-validated SIC algorithms, with two main differences: additional manual quality control and inclusion of the period from 1978–1987, for which SIC data are available every other day^{54,55}. From January 2018 to February 2019, we use the NOAA/NSIDC Near-Real-Time (NRT) CDR v1 product⁵⁶. The combined ‘NSIDC Merged’ record from 1978–2019 was used for SIC climatology (Fig. 2a) and polynya identification (Fig. 5a). Nimbus-5 ESMR was included in the polynya record to highlight the 1974–1976 events, although precise quantitative comparison with NSIDC Merged is not possible. For more qualitative analyses, we use higher-resolution daily SIC data from the Advanced Microwave Scanning Radiometer (AMSR) sensors (see Fig. 2a for comparison with NSIDC Merged). We use the University of Bremen v5 AMSR-E (2002–2011) and AMSR2 (2012–2019) products, derived using the ARTIST Sea Ice (ASI) algorithm at a resolution of 6.25 km^{57,58}.

SIC climatology (Fig. 2a) includes days with valid data from at least 75% of grid cells in the Maud Rise region. Median and 25%–75% IQR time series were generated by compositing over day-of-year, then filtering with a 7-day centred running mean. Along-trajectory SIC (Fig. 4) represents the average SIC from AMSR2-ASI (from 4 July 2012 onwards) or NSIDC Merged (from 19 December 2011 to 3 July 2012) within a moving 1° latitude × 2° longitude box centred on the given or estimated location for each float profile.

Polynya identification. A polynya is defined as any nonlinear-shaped opening within sea ice that contains open water, brash ice, new ice, nilas and/or young ice⁵⁹. Passive microwave sensing generally underestimates SIC for new, thin ice, aiding the detection of polynyas^{60,61}. Following past studies identifying polynyas in the Cosmonaut Sea^{62,63}, we define ‘polynya extent’ as the sum of connected pixels (allowing diagonal connections) with SIC beneath some threshold. We excluded areas connected to the open ocean (that is, embayments), common during transitional freeze and melt periods, and calculated the centroid of individual openings. Patches of ice within connected open regions were negated using a binary dilation algorithm to ensure that polynya extent, rather than area, was quantified.

Figure 5a highlights midwinter openings that appear from closed ice cover, as occurred in 2016 and 2017 (Extended Data Figs. 2, 8). These differ from early-winter embayments that later became enclosed, such as occurred in 1974 (ref. ²⁴). The latter probably reflects ocean preconditioning strong enough to permit overturning from slight amounts of ice growth or even cooling alone (a ‘thermal mode’ of stratification)^{2,35,41,64}. This mode does not require perturbations such as storms, which we identify as critical for midwinter openings. For comparison, we included the 1974–1976 polynyas in our record despite their different mode of formation.

We consider the period of closed ice cover over Maud Rise to be delimited each winter by one week after the first date of 90% average SIC and one week before the last date of 90% SIC, consistent with climatological SIC²⁵. The one-week buffers account for gradual ice advance and retreat, during which small, brief openings frequently form via enclosure and climatological melt²³. Fixed start and/or end dates of 1 July and 31 October, consistent with climatology (Fig. 2a), were used for freeze and/or melt seasons with already-established polynyas (1974–1976, 2017) or some gaps in SIC data (1973). Other winters with major data gaps (1972, 1977, 1978, 1987) were omitted from the record. The summed daily extent of polynyas with centroids in the eastern Weddell region was calculated using SIC thresholds

of 40%, 50% and 60% for comparison, because grid connectivity is sensitive to SIC threshold. These thresholds reflect the dynamic mix of open water and new ice within polynyas⁶¹, and lie between the low (for example, 15%) thresholds used to define the seasonal ice edge and high (for example, 80%) thresholds previously used to identify Antarctic offshore polynyas^{62,63}. Dates with non-zero total polynya extent at the 60% threshold, chosen to maximally differentiate years from one another, were designated ‘polynya days’. As an approximation, polynya day counts from 1979–1986 were doubled because SIC data are available every other day. Our average SIC threshold of 50% was used in Fig. 1a, which depicts the largest polynya identified within the Maud Rise region on each date, and Figs. 2a, 3a, which track the summed extent of polynyas with centroids in the eastern Weddell region.

The accuracy of our methodology was checked by visually inspecting SIC images. Our assessment is consistent with previous reports of Maud Rise polynyas in 1973^{2,23}, 1980²³, 1994^{30,51,65} and 2005³¹. A previous analysis for the Maud Rise region of total days with average SIC under 92% from 1979–2004⁴⁷ is qualitatively similar to our polynya record. Given the association of deep convection with coherent openings¹, however, we focus on discrete polynya events, rather than slight reductions in SIC. We note that a small opening outside the detection parameters of our algorithm (and therefore absent in Fig. 5a) appeared above Maud Rise in late October 2018, just before climatological melt.

Hydrographic data. Argo profiling floats drift at a depth of around 1,000 m and sample every 7–10 days, profiling on ascent⁶⁶. Sampling intervals are generally 2 dbar above 1,000 m and 100 dbar below 1,000 m (ref. ⁶⁷). Measurements are rated to accuracies of at least 0.005°C, 0.01 psu and 2.5 dbar (ref. ⁶⁶). An ice-avoidance algorithm monitors median temperature between 50 and 20 m, aborting ascent if near the freezing point, which indicates ice cover^{67–69}. Under-ice profiles are stored without position fixes and transmitted upon spring ice melt. A rare exception is when a float surfaces within a polynya, as occurred in 2017 with SOCCOM floats 5904468 and 5904471 (Extended Data Fig. 8). Because the algorithm requires three consecutive ice-free determinations⁶⁷, both floats inferred warm conditions in or near the polynya over about 21 days before surfacing.

All Argo temperature and salinity measurements (Extended Data Fig. 1) south of 55° S and between 70° W and 50° E through 1 October 2018 were downloaded from the US-GODAE Global Data Assembly Center (GDAC)⁷⁰. We rejected profiles without timestamps, consecutive under-ice profiles with timestamps erroneously reflecting transmission upon ice melt, and data from floats affected by position and date jumps and other quality control issues.

Under-ice or missing positions (quality control flags ‘8’ or ‘9’) were linearly interpolated, using profile timestamps to assign distances along great circle routes between known GPS fixes. Advanced methods, such as interpolation along contours of planetary potential vorticity, could reduce position uncertainty over some areas of the Weddell Sea; however, this may not be the case near Maud Rise, particularly to the southwest and northeast of the seamount⁷¹. Given the tendency of floats to remain trapped over Maud Rise (Fig. 1b, c) owing to Taylor column dynamics^{31,51,72}, it is unlikely that floats 5903616, 5904468 and 5904471 deviated away from the seamount during the winters of 2012 through to 2017. SOCCOM float deployments were conducted with this phenomenon in mind⁷³.

Both real-time (‘R’) and delayed-mode (‘D’) data were obtained. Although both have passed automatic quality control checks, only delayed-mode data have undergone detailed quality control inspection, although sensor drift calibrations often extend automatically to real-time profiles⁷⁰. If a profile was available on delayed-mode, we used its adjusted parameters. The first cast within a profile file was selected if multiple were available. Only depths at which all three quality control flags for pressure, temperature, and salinity were good (‘1’) or probably good (‘2’) were extracted.

Float 5904468 represents an exception to this procedure. Drift in its salinity sensor, a Sea-Bird SBE-41CP conductivity cell, began in early 2016 at a rate of approximately 0.07 psu per year, constant with depth and linear in time. This drift was corrected through 13 May 2017 (profile 83) by the GDAC using standard procedures (weighted least-squares fit to deep climatology based on nearby Argo data). We extended this correction from 23 May 2017 (profile 84) to 8 May 2018 (profile 118) by subtracting the linear trend of 1,600–1,700 m average salinity, calculated from profiles linearly interpolated to 1-m spacing. Corrected 1,600–1,700 m average salinity agrees with that measured by the nearby float 5904471 over 2015–2018 (see Fig. 1b, c) to within about 0.01 psu except for periods of mixing during the 2017 polynya. Profiles from 119 onwards were not used.

Salinity measurements are provided by the GDAC with a precision of 0.001 psu. The combination of weak vertical salinity gradient in the Weddell Sea and fine sampling interval (2 dbar) creates artificial 0.001 psu steps, generating spurious static instabilities. We applied a mild quadratic smoothing spline (*UnivariateSpline* from SciPy with smoothing factor $s = 0.00015$) to all float salinity profiles, attenuating these jumps while preserving integrity of the profile. This procedure essentially downsamples the salinity measurements to allow buoyancy frequency to be accurately resolved at all depths.

Dissolved oxygen measurements from floats 5904468 and 5904471 were downloaded from the SOCCOM quality-controlled archive⁷⁴. Dissolved oxygen data from the non-SOCCOM float 5903616 were obtained from the University of Washington (UW) Calibrated O₂ package, v1.1⁷⁵, with updated profiles provided by R. Drucker (personal communication). The sampling interval of dissolved oxygen is coarser than for CTD data, ranging from about 5 dbar above 100 m to about 100 dbar below 1,000 m⁶⁷. Dissolved oxygen optodes onboard most SOCCOM floats collect in-air samples for direct calibration, allowing for accuracy of about 1% of near-surface values (around 325 $\mu\text{mol kg}^{-1}$), or 3 $\mu\text{mol kg}^{-1}$ (ref. ⁷⁶). The optode on 5903616 lacks this capability and instead has been calibrated to deep reference data, with an accuracy of about 2%, or 7 $\mu\text{mol kg}^{-1}$ (ref. ⁷⁵). Comparison of UW-calibrated dissolved oxygen from 5903616 with SOCCOM-calibrated dissolved oxygen from 5904468 and 5904471—performed in surface-referenced potential density space—revealed a positive bias of about 6.0 $\mu\text{mol kg}^{-1}$, uniform in depth, during their overlap period at Maud Rise (2015–2016). We subtracted this offset from all UW-O₂ profiles by float 5903616.

Shipboard CTD and bottle data and instrumented elephant seal profiles (Extended Data Fig. 1) featuring both temperature and salinity were obtained at original depth levels from the World Ocean Database 2018 prerelease⁷⁷ with August 2018 additions. Depth levels were retained only if global quality control flags for depth, temperature and salinity profiles as well as the three quality control flags at that depth were marked good ('0').

Instrumented elephant seals typically dive to about 600 m while foraging, although dives to 2,000 m occasionally occur⁷⁸. CTD measurements begin at the deepest point of a dive and continue during ascent; data may be collected on as many as four dives per day⁷⁹. Owing to data transfer and energy constraints, casts are transmitted in compressed form. A 'broken-stick' algorithm selects 10–25 depths that best represent the profile; thus, resolution may be coarse for deeper profiles^{78,79}. The filters we apply when creating hydrographic climatologies account for this possibility (see below). Measurement accuracy is estimated at 0.04 °C and 0.03 psu⁸⁰, lower than for floats; satellite fixes are accurate to about 5 km (ref. ⁷⁸). **Derived oceanographic quantities.** The Python implementation of the Gibbs SeaWater Oceanographic Toolbox of TEOS-10 (<https://teos-10.github.io/GSW-Python/>) was used to compute profiles of depth or pressure, potential temperature, surface-referenced potential density, and buoyancy (Brunt–Väisälä) frequency squared (N^2). We interpolated N^2 profiles to 1-m spacing, then applied a 50-m centred running mean. This filtering reduces noise and spikes, allowing patches of consistently low N^2 to be visible in Fig. 4d and Extended Data Fig. 6d despite the compressed vertical axis.

Mixed-layer depths (MLD; Fig. 4, Extended Data Figs. 3c, 6) were determined as the depth at which surface-referenced potential density exceeds its value at $z_{\text{ref}} = 10$ m, estimated using interpolation or nearest-neighbour extrapolation, by a threshold of $\Delta\sigma_\theta = 0.03 \text{ kg m}^{-3}$ (refs ^{81,82}). Mixed-layer temperature (MLT) and mixed-layer salinity (MLS) were averaged from profiles interpolated to 0.1-m spacing and extrapolated to the surface. We note that MLDs sometimes shoaled sharply during the 2016 and 2017 polynyas, rather than deepening (Extended Data Fig. 3c). This reflects the development of shallow fresh layers from ice melt (Fig. 2b), as seen in idealized ice–ocean model experiments with storm perturbations³³. The two profiling floats, sampling coarsely in time and space, did not observe active overturning, which probably occurred within plumes of horizontal scale $O(100 \text{ m})$ occupying a fraction of the convection region^{40,83,84} at locations with a weaker melt layer or pycnocline. The observed interior mixed layers (Extended Data Fig. 6) may have been formed by nearby deep convection, communicated laterally through efficient mixing due to baroclinic instability.

Upper-ocean freshwater anomaly (Extended Data Fig. 3e), or 'salt deficit'^{33,34}, was integrated from 0–250 m, encompassing MLDs in all seasons (see Extended Data Fig. 3c) and salinity variability an order of magnitude greater than below 250 m. The metric was computed as in ref. ³³, in terms of the freshwater provided by melt of an equivalent sea ice thickness (units of metres):

$$\eta(250 \text{ m}) = \frac{1}{\Delta S_i} \int_{0 \text{ m}}^{250 \text{ m}} [S(250 \text{ m}) - S(z)] dz$$

where S is salinity and ΔS_i is the approximate salinity difference between ocean and sea ice (about 30 psu)³⁴. Trapezoidal integration was applied to profiles interpolated and extrapolated as above.

To interpret the climatological 0–250 m freshwater anomaly (grey shading in Extended Data Fig. 3e), we assume that brine rejection approximately balances ice melt from local and nonlocal sources, because Maud Rise is near the circumpolar line of zero annual net freshwater flux associated with sea ice⁸⁵, and that salt fluxes from geostrophic advection are relatively small. This implies that positive salt fluxes from Ekman upwelling (due to cyclonic wind stress curl; see Fig. 5e) approximately balance precipitation and evaporation on an annual net basis.

We thus interpret the seasonal cycle of climatological 0–250 m freshwater anomaly as predominantly reflecting brine rejection and ice melt. A similar approach has been used to estimate ice formation rates from instrumented seal profiles offshore of East Antarctica⁸⁶.

Convection resistance (CR) represents the buoyancy loss required for overturning to reach a given depth, H :

$$\text{CR}(H) = \frac{g}{\rho_0} \int_{0 \text{ m}}^H [\sigma_\theta(H) - \sigma_\theta(z)] dz$$

where g is acceleration due to gravity, ρ_0 is a seawater reference density (1,027.8 kg m^{-3}), and σ_θ is surface-referenced potential density^{19,87,88}. Depth sections of convection resistance illustrate that low interior stratification combined with a weakened winter halocline, as occurred in 2016 and 2017, may permit deep-reaching convection near Maud Rise (Extended Data Fig. 6e).

Composites of float time series. Data from floats 5903616 (active only until 2 June 2016), 5904468 and 5904471 were combined to create time series of MLS, MLD, MLT and 0–250 m freshwater anomaly in 2016 and 2017 (Fig. 2b, Extended Data Fig. 3c–e). Float measurements were linearly interpolated to daily resolution, and the daily means (for MLT and MLD), maxima (for MLS), or minima (for freshwater anomaly) were calculated. MLS and MLT agree between the three floats to within 0.05 psu and 0.25 °C, except in the 2–3 months after ice melt. Freshwater anomaly and MLD are more variable, but fluctuations generally occur synchronously. Maximum MLS and minimum freshwater anomaly were chosen to highlight the most extreme preconditioning observed near Maud Rise, motivated by the notion that overturning will preferentially occur where stratification is weakest. Given the substantial spatial inhomogeneity around Maud Rise^{31,32}, sampling by 2–3 floats probably underestimates the most extreme preconditioning that occurred in 2016–2017.

The red line in Fig. 2b represents the MLS required to eliminate stratification between the mixed layer and 250 m, at which point initiation of deep convection would be trivial³. This 'overturning limit' is fresher than salinity at 250 m owing to the destabilizing warmth of the thermocline³³. Its value was determined as the MLS at which surface density, computed using composite (average) MLT, would exceed composite surface-referenced potential density at 250 m. The use of surface density neglects the development of thermobaric instabilities, which could hasten overturning near this limit of zero stratification³⁶.

Salinity fluxes from upwelling. Upwelling due to divergence of Ekman transport can cause heat and salt fluxes into the mixed layer^{28,35,89}. Vertical velocity w_{Ek} at the base of the Ekman layer is related to wind stress curl⁹⁰ in the absence of ice cover:

$$w_{\text{Ek}} = \nabla \cdot \mathbf{U}_{\text{Ek}} = \mathbf{k} \cdot \nabla \times \frac{\boldsymbol{\tau}}{\rho_0 f}$$

where \mathbf{U}_{Ek} is the horizontal vector Ekman transport, \mathbf{k} is the vertical unit vector, $\boldsymbol{\tau}$ is the vector surface wind stress, ρ_0 is a seawater reference density (1,027.8 kg m^{-3}), and f is the Coriolis parameter.

Upwelling prevails over Maud Rise because wind stress curl is consistently negative (Fig. 5e). We estimated, in back-of-the-envelope fashion, the anomalous mixed-layer salt flux resulting from record cyclonic wind stress curl over Maud Rise in the ice-free months of 2015 and 2016—before the 2016 polynya—compared to a typical previous year (Fig. 5e; Methods section 'Atmospheric reanalysis'). The average wind stress curl from ERA-I over 1979–2014 was $-2.06 \times 10^{-7} \text{ N m}^{-3}$, whereas the averages for January–May 2015 and 2016 were $-2.90 \times 10^{-7} \text{ N m}^{-3}$ and $-3.08 \times 10^{-7} \text{ N m}^{-3}$, respectively, increasing upwelling by 8.0 m and 9.8 m.

To estimate the corresponding mixed-layer salt flux, salinity gradients across the base of the mixed layer were measured using observations from floats 5903616, 5904468 and 5904471 over January–May 2015 and 2016. MLD during these periods averaged 41.3 m and 56.2 m, respectively. MLS averaged 33.90 psu and 34.22 psu, while the average salinities of depth levels between the time-varying MLD and 7.8 m or 9.6 m below the MLD were 34.09 psu and 34.34 psu, respectively. Expected MLS elevation was computed using these values:

$$\Delta S_{\text{ml}} = \frac{h_{\text{up}} S_{\text{up}} + (h_{\text{ml}} - h_{\text{up}}) S_{\text{ml}}}{h_{\text{ml}}} - S_{\text{ml}}$$

where S_{ml} and S_{up} are MLS and sub-mixed-layer salinity, and h_{ml} and h_{up} represent MLD and thickness of the upwelled layer. We determine that additional Ekman upwelling elevated MLS by at least 0.04 psu in 2015 and 0.02 psu in 2016, for a total of at least 0.06 psu. These are lower bounds because upwelling continues in winter, although the absence of ice–ocean stress data precludes the estimation of upwelling during ice-covered months. Statistically propagating the Argo salinity measurement accuracy of ± 0.01 psu yields a ΔS_{ml} uncertainty of just ± 0.002 psu; however, additional contributions from the exclusion of ice-covered months,

coarseness of sampling, spatial inhomogeneity and time averaging are less easily quantified.

Hydrographic climatologies. Float, shipboard and instrumented seal measurements were used to create climatologies of average MLS, MLD, MLT and 0–250 m freshwater anomaly near Maud Rise ($R < 250$ km) and away from Maud Rise ($250 \text{ km} < R < 500$ km; Fig. 2b, Extended Data Fig. 3c–e). Profiles with a shallowest measurement below 30 m, deepest measurement above 250 m, or MLD greater than 250 m were excluded, ensuring that the mixed layer and upper seasonal pycnocline were captured. These filters yielded 1,035 float profiles from 50 floats, 265 shipboard casts, and 124 seal profiles near Maud Rise, and 1,523 float profiles from 79 floats, 510 shipboard casts, and 445 seal profiles away from Maud Rise (Extended Data Fig. 1). Mean and median years of the compiled observations are 2008 and 2012 for Maud Rise and 2006 and 2008 away from Maud Rise. Fewer than 1% of these observations were collected before 1970.

Calculated metrics were composited by day-of-year within 365 overlapping 21-day bins, looping from December to January. Median and 25%–75% IQR were computed for each bin instead of mean and standard deviation to reduce sensitivity to outliers and reflect skewness in spatial and temporal variability. Lastly, a 21-day centred running mean was applied. We note that the absence of abrupt changes in these metrics—for example, sharp increases in MLS during ice formation—is an artefact of the compositing method. Years featuring early and late ice formation are combined, for example, resulting in gradual changes.

Sub-pycnocline temperature records. Extended Data Fig. 9d, e depicts 15.5 years (2002–2017) of 258-m temperature anomalies from climatology within the eastern Weddell region ($R < 500$ km from Maud Rise) based on 2,445 float, 151 shipboard and 407 seal profiles, and 250–1,000 m average temperature anomalies based on 2,421 float and 133 shipboard profiles and one seal profile. We compare observations with a gridded $1/4^\circ$ monthly climatology, the 2018 WOCE/Argo Global Hydrographic Climatology⁹¹ (WAGHC).

We included profiles with a shallowest measurement above 30 m and a deepest measurement below the depth(s) of interest. Temperatures at 258 m or at the 14 WAGHC levels from 218–1,050 m were estimated from each profile using linear interpolation. Observations were co-located in space with WAGHC fields using nearest-neighbour interpolation, then the climatological value on a given day-of-year was estimated using linear interpolation over values assigned to the 15th of each month. The average 250–1,000 m anomaly was computed after linearly interpolating anomalies at the 14 depths (218–1,050 m) to 1-m spacing. Lastly, the anomalies were binned biannually, with at least five samples required in each bin. Extended Data Fig. 9d, e displays median and 25%–75% IQR values in addition to violin plots, which represent a kernel density distribution of the binned data.

Decadal changes in Weddell gyre heat content are probably closely related to Circumpolar Deep Water inflow variability^{92,93}, but uncertainty remains regarding the importance of internal processes such as offshore polynyas^{94,95}. Ventilation during the 1974–1976 openings cooled depths from 200–2,700 m by 0.4°C within the polynya area⁹⁶, a signal that propagated into deep and abyssal waters away from the Weddell Sea through advective transport and wave mechanisms^{4,97}. Nonetheless, local rebound in heat content near Maud Rise had occurred by 1984⁹⁴. The absence of polynya activity in the latter half of the 1980s (Fig. 5a) is thus further evidence against a recurrence timescale governed exclusively by deep heat recharge.

Polynya heat flux estimates. Gradual deep cooling and freshening observed by both floats from 2015 to early 2016 (Extended Data Fig. 6a, b) probably reflects movement of floats from the warm ‘halo’ encircling Maud Rise into its overlying Taylor column regime (Fig. 1b). However, subsequent deep anomalies are inconsistent with hydrographic surveys conducted in 1994⁵¹ and 2005⁵¹, despite conditions coincidentally also favouring polynya formation in those years (Fig. 5a). We thus attribute these abrupt decreases in heat content primarily to ventilation during the 2016 polynya. We difference sub-mixed-layer potential temperature profiles for each float immediately before 27 July 2016 with those over the five succeeding months until 1 January 2017 to estimate the rate of convective heat loss (Q , in W m^{-2}):

$$Q = \frac{\rho_0 c_p}{\Delta t} \int_{1,650 \text{ m}}^{200 \text{ m}} \Delta\theta(z) dz$$

where ρ_0 is a seawater reference density ($1,027.8 \text{ kg m}^{-3}$), c_p is the specific heat of seawater ($3,850 \text{ J kg}^{-1} \text{ }^\circ\text{C}^{-1}$), Δt is the polynya duration (21 days), and $\Delta\theta(z)$ is the change in potential temperature at depth z . The median heat flux is 208 W m^{-2} from 31 profiles, 74% of which yield estimates between 0 W m^{-2} and the average ocean–atmosphere turbulent heat flux during the polynya, 252 W m^{-2} (Fig. 3c, Extended Data Fig. 7). Estimates outside this range may reflect lateral mixing or float displacement.

Southern Annular Mode. In its positive phase, SAM is characterized by lower pressure over Antarctica than over mid-latitudes, associated with poleward

displacement of the mid-latitude westerly jet⁴⁴. We use a monthly index of SAM from 1972–2019⁹⁸. The index is filtered using a two-year centred running mean in Fig. 5b to highlight lower-frequency fluctuations.

Atmospheric reanalysis. Six-hourly and monthly-mean 0.75° -resolution surface-level analysis and forecast fields from ERA-Interim (ERA-I), a third-generation global reanalysis product⁹⁹, were obtained from January 1979 to December 2018. Forecast output was de-accumulated, and climatologies were constructed by compositing over day-of-year, calculating mean and 25%–75% IQR values, and filtering using a seven-day centred running mean.

10-m wind speed (Fig. 3b, Extended Data Fig. 4) was calculated as:

$$|u| = \sqrt{u^2 + v^2}$$

Wind stress curl (Fig. 5e) was computed using second-order-accurate central differencing:

$$\nabla \times \tau = \frac{\Delta\tau_y}{\Delta x} - \frac{\Delta\tau_x}{\Delta y}$$

Given the uncertainty of ice motion products near ice–ocean boundaries and in areas of low SIC¹⁰⁰, we instead estimate wind-forced ice drift velocities (u_i and v_i) from ERA-I using a ‘rule-of-thumb’ for thin Weddell Sea pack ice that suggests drift is around 3% of wind speed at a turning angle of around 23° to the left of winds¹⁰¹. Others have used a scaling of 2% and turning angle of 30° (ref. ¹⁰²); our results are not sensitive to the choice of these parameters. Potential ice divergence due to wind forcing (Extended Data Fig. 5) was inferred from these drift velocities:

$$\nabla \cdot u_i = \frac{\Delta u_i}{\Delta x} + \frac{\Delta v_i}{\Delta y}$$

Because geostrophic winds are nondivergent, the response of a spatially uniform ice cover will also be nondivergent despite the turning angle¹⁰³. Ice divergence therefore requires deviations from geostrophy, such as sub-geostrophic speeds associated with gradient winds in an intense cyclone, or spatially inhomogeneous ice properties. Extended Data Fig. 5 highlights potential ice divergence associated with the former. Divergence during storm episodes probably also occurs owing to ice deformation on shorter spatial scales than those resolved by ERA-I³⁷.

Under weak winds, ocean heat loss within a polynya may warm the lower atmosphere enough to generate a persistent low-pressure anomaly, cyclonic wind field, and ice divergence, thus maintaining the opening^{2,35}. This positive feedback is not apparent in ERA-I in 2016 (Extended Data Fig. 5) or 2017 (not shown), perhaps because anomalies are quickly advected downstream by background winds and transient storms.

Ocean–atmosphere turbulent heat fluxes were estimated from ERA-I fields using the COARE 2.0 bulk flux algorithm¹⁰⁴, as described in ref. ¹⁰⁵. Calculated open-water heat fluxes for each grid cell were scaled by open water fraction ($1 - \text{SIC}$) from ERA-I and averaged within the Maud Rise region. Dividing by the mean open water fraction yields the average heat flux within the 2016 polynya (Fig. 3c).

Weddell Low. We designate the Weddell Low as the climatological low mean sea-level pressure (MSLP) centre over the Riiser-Larsen Sea, around which a large-scale geostrophic wind field circulates¹⁰⁶. The curl of this wind field has a leading role in controlling the strength of the Weddell gyre^{107–109}. By analogy to the Amundsen Sea Low⁴⁵, the Weddell Low may be expected to modulate cyclone central pressures and spatial density across the Weddell Sea region. We characterize the Weddell Low as the region $\pm 1\sigma$ in latitude and longitude around the minimum ERA-I monthly mean MSLP identified within $60^\circ\text{--}70^\circ\text{S}$, $60^\circ\text{W--}60^\circ\text{E}$. The average minimum occurs at 64.4°S , 16.7°E , resulting in an area bounded by $61.9^\circ\text{--}66.8^\circ\text{S}$, $12.5^\circ\text{W--}45.8^\circ\text{E}$. The Weddell Low index represents monthly mean MSLP within this region, filtered using a two-year centred running mean as above (Fig. 5c).

Meteorological station records. To complement and extend the reanalysis record, Fig. 5c also includes MSLP data from Novolazarevskaya Station on the Queen Maud Land coast of Antarctica (71°S , 12°E ; location marked in Fig. 1a; data available from 1972–2019). Monthly means were obtained from the Reference Antarctic Data for Environmental Research (READER) archive¹¹⁰. For comparison, Extended Data Fig. 9b shows monthly mean MSLP from the nearby Neumayer (71°S , 8°W ; 1981–2019) and Syowa (69°S , 40°E ; 1957–2019) Stations from the READER archive and from Maitri Station (71°S , 12°E ; 1990–2019), South African National Antarctic Expedition Station (SANAE; 70°S , 2°W ; 1973–1994), and SANAE-Automatic Weather Station (SANAE-AWS; 72°S , 3°W ; 1997–2019), obtained as subdaily records from the NOAA-NCEI Integrated Surface Database (ISD)¹¹¹. The first two years of the Neumayer and SANAE-AWS records were dropped owing to sparse data and quality issues. As above, two-year centred running means are presented. We observe consistent fluctuations between the station records. Strong correlation between Novolazarevskaya and the Weddell Low

($r = 0.84$; see Extended Data Table 1) suggests that the former is a useful proxy for the latter during the pre-ERA-I period.

Storm identification. A high degree of similarity between six-hourly MSLP from SNAE-AWS, south of Maud Rise, and the nearest ERA-I grid cell from 1997–2019 ($r = 0.93$; mean absolute deviation = 2.2 hPa; mean bias = 0.8 hPa) suggests that ERA-I skilfully represents variability near Maud Rise associated with synoptic-scale weather systems. Validation with independent pressure data from drifting ice buoys in the Bellingshausen Sea also suggests that ERA-I is skilled at capturing individual storms in the Southern Ocean, particularly relative to other available reanalyses¹¹². We identify the passage of particularly intense cyclones within the Maud Rise (Figs. 2a, 3a) and eastern Weddell (Fig. 5d) regions. Previous work indicates that storms with a central pressure below 950 hPa represent the most extreme 5% of individually tracked cyclones at a latitude of 65°S¹¹³. Adopting this pressure threshold, we find (coincidentally) that 5.3% of ERA-I six-hourly MSLP fields feature one or more grid cells below 950 hPa within the eastern Weddell region. Statistically, this corresponds to a wind speed threshold of about 20 m s⁻¹, which is exceeded in 5.8% of instantaneous six-hourly fields. Days on which either criteria are met are designated ‘storm days’ (see Extended Data Fig. 4). These are denoted by vertical bars in Figs. 2a, 3a, Extended Data Fig. 3a and counted annually during ice-covered winter (May to October) in Fig. 5d. In plots spanning 2016 and 2017, sequences of storm days separated by a single non-storm day are aggregated for ease of visualization.

Data availability

The data analysed in this article are all publicly available, with the exception of updates to the UW Calibrated O₂ package, described below:

Ocean bathymetry data were obtained from the ETOPO1 1 Arc-Minute Global Relief Model¹¹⁴ at <https://doi.org/10.7289/V5C8276M> (accessed February 2017).

Sea ice concentration data were obtained for the period 1972–1977 from the NSIDC Nimbus-5 ESMR v1 product⁵³ at <https://doi.org/10.5067/W2PKTWMTY0TP> (accessed February 2017); for the period 1978–2017 from the merged NASA Goddard v3 product⁵⁴ at <https://doi.org/10.7265/N59P2ZTG> (accessed October 2018); for January 2018–February 2019 from the NOAA/NSIDC Near-Real-Time CDR v1 product⁵⁶ at <https://doi.org/10.7265/N5FF3QJ6> (accessed February 2019); and for the period 2002–2019 from the University of Bremen ASI AMSR-E and AMSR2 v5 products^{57,58} at <https://seaiice.uni-bremen.de/sea-ice-concentration/> (accessed February 2019).

Profiling float temperature and salinity measurements were obtained from the US-GODAE GDAC⁷⁰ at <http://www.usgodae.org/ftp/outgoing/argo> (accessed October 2018). Dissolved oxygen measurements for floats 5904468 and 5904471 were obtained from the SOCCOM quality-controlled archive⁷⁴ at <https://doi.org/10.6075/J02J6968> (accessed January 2019) and for float 5903616 from the UW Calibrated O₂ package, v1.1⁷⁵ at <http://runt.ocean.washington.edu/o2> (accessed August 2016), with updated dissolved oxygen profiles provided by R. Drucker (personal communication, August 2017).

Shipboard and elephant seal temperature and salinity measurements were obtained from the World Ocean Database 2018 prerelease⁷⁷ with August 2018 additions at <http://www.nodc.noaa.gov/OC5/SELECT/dbsearch/dbsearch.html> (accessed October 2018).

Gridded climatological ocean temperature fields were obtained from the 2018 WAGHC⁹¹ at <http://icdc.cen.uni-hamburg.de/1/daten/ocean/waghc> (accessed January 2018).

The monthly SAM index⁹⁸ was obtained for the period 1972–2019 at <http://legacy.bas.ac.uk/met/gjma/sam.html> (accessed February 2019).

Monthly and daily ERA-I atmospheric reanalysis fields⁹⁹ were obtained for the period 1979–2018 using the Python MARS API, described at <https://confluence.ecmwf.int/display/WEBAPI/> (accessed February 2019).

Queen Maud Land pressure records (see Methods section ‘Meteorological station records’) were obtained from the READER archive¹¹⁰ at <http://legacy.bas.ac.uk/met/READER> (accessed February 2019) and the NOAA-NCEI ISD¹¹¹ at <http://www.ncdc.noaa.gov/isd> (accessed February 2019).

Code availability

Analytical scripts used to generate the figures in this paper are available at <https://github.com/ethan-campbell>.

51. Muench, R. D. et al. Maud Rise revisited. *J. Geophys. Res.* **106**, 2423–2440 (2001).
52. Meier, W. N., Gallaher, D. & Campbell, G. G. New estimates of Arctic and Antarctic sea ice extent during September 1964 from recovered Nimbus I satellite imagery. *Cryosphere* **7**, 699–705 (2013).
53. Parkinson, C. L., Comiso, J. C. & Zwally, H. J. *Nimbus-5 ESMR Polar Gridded Sea Ice Concentrations* v.1 <https://doi.org/10.5067/W2PKTWMTY0TP> (National Snow and Ice Data Center, 2004).
54. Meier, W. N. et al. *NOAA/NSIDC Climate Data Record of Passive Microwave Sea Ice Concentration* v.3 <https://doi.org/10.7265/N59P2ZTG> (National Snow and Ice Data Center, 2017).
55. Meier, W. N., Peng, G., Scott, D. J. & Savoie, M. H. Verification of a new NOAA/NSIDC passive microwave sea-ice concentration climate record. *Polar Res.* **33**, <https://doi.org/10.3402/polar.v33.21004> (2014).
56. Meier, W. N., Fetterer, F. & Windnagel, A. K. *Near-Real-Time NOAA/NSIDC Climate Data Record of Passive Microwave Sea Ice Concentration* v.1 <https://doi.org/10.7265/N5FF3QJ6> (National Snow and Ice Data Center, 2017).
57. Spreen, G., Kaleschke, L. & Heygster, G. Sea ice remote sensing using AMSR-E 89-GHz channels. *J. Geophys. Res. Oceans* **113**, 1–14 (2008).
58. Beitsch, A., Kaleschke, L. & Kern, S. Investigating high-resolution AMSR2 sea ice concentrations during the February 2013 fracture event in the Beaufort Sea. *Remote Sens.* **6**, 3841–3856 (2014).
59. JCOMM Expert Team on Sea Ice. *Sea-Ice Nomenclature*. WMO No. 259 (World Meteorological Organization, 2014).
60. Comiso, J. C., Cavalieri, D. J., Parkinson, C. L. & Gloersen, P. Passive microwave algorithms for sea ice concentration: a comparison of two techniques. *Remote Sens. Environ.* **60**, 357–384 (1997).
61. Comiso, J. C. & Steffen, K. Studies of Antarctic sea ice concentrations from satellite data and their applications. *J. Geophys. Res. Oceans* **106**, 31361–31385 (2001).
62. Comiso, J. C. & Gordon, A. L. Cosmonaut polynya in the Southern Ocean: structure and variability. *J. Geophys. Res. Oceans* **101**, 18297–18313 (1996).
63. Arbeter, T. E., Lynch, A. H. & Bailey, D. A. Relationship between synoptic forcing and polynya formation in the Cosmonaut Sea: 1. Polynya climatology. *J. Geophys. Res.* **109**, C04022 (2004).
64. Gordon, A. L. in *Elsevier Oceanography Series: Deep Convection and Deep Water Formation in the Oceans* Vol. 57 (eds. Chu, P. C. & Gascard, J.-C.) 17–35 (Elsevier, 1991).
65. Venegas, S. A. & Drinkwater, M. R. Sea ice, atmosphere and upper ocean variability in the Weddell Sea, Antarctica. *J. Geophys. Res.* **106**, 16747–16765 (2001).
66. Riser, S. C. et al. Fifteen years of ocean observations with the global Argo array. *Nat. Clim. Change* **6**, 145–153 (2016).
67. Riser, S. C., Swift, D. & Drucker, R. Profiling floats in SOCCOM: technical capabilities for studying the Southern Ocean. *J. Geophys. Res. Oceans* **123**, 4055–4073 (2018).
68. Klatt, O., Boebel, O. & Fahrbach, E. A profiling float’s sense of ice. *J. Atmos. Ocean. Technol.* **24**, 1301–1308 (2007).
69. Wong, A. P. S. & Riser, S. C. Profiling float observations of the upper ocean under sea ice off the Wilkes Land coast of Antarctica. *J. Phys. Oceanogr.* **41**, 1102–1115 (2011).
70. Carval, T. et al. *Argo User’s Manual* v. 3.2 (Argo, 2017).
71. Chamberlain, P. M. et al. Observing the ice-covered Weddell Gyre with profiling floats: position uncertainties and correlation statistics. *J. Geophys. Res. Oceans* **123**, 8383–8410 (2018).
72. Meredith, M. P. et al. Circulation, retention, and mixing of waters within the Weddell-Scotia Confluence, Southern Ocean: The role of stratified Taylor columns. *J. Geophys. Res. Oceans* **120**, 547–562 (2015).
73. Talley, L. D. et al. Southern Ocean biogeochemical float deployment strategy, with example from the Greenwich Meridian line (GO-SHIP A12). *J. Geophys. Res. Oceans* **124**, 403–431 (2019).
74. Johnson, K. S. et al. *Southern Ocean Carbon and Climate Observations and Modeling (SOCCOM) Float Data Archive - Snapshot 2018-12-31*, <https://doi.org/10.6075/J02J6968> (UC San Diego, 2019).
75. Drucker, R. & Riser, S. C. In situ phase-domain calibration of oxygen Optodes on profiling floats. *Methods Oceanogr.* **17**, 296–318 (2016).
76. Johnson, K. S. et al. Biogeochemical sensor performance in the SOCCOM profiling float array. *J. Geophys. Res. Oceans* **122**, 6416–6436 (2017).
77. Boyer, T. P. et al. *World Ocean Database 2018*. NOAA Atlas NESDIS 87 (NOAA, 2018).
78. Roquet, F. et al. A Southern Indian Ocean database of hydrographic profiles obtained with instrumented elephant seals. *Sci. Data* **1**, 140028 (2014).
79. Boehme, L. et al. Animal-borne CTD-Satellite Relay Data Loggers for real-time oceanographic data collection. *Ocean Sci.* **5**, 685–695 (2009).
80. Siegelman, L. et al. Correction and accuracy of high- and low-resolution CTD data from animal-borne instruments. *J. Atmos. Ocean. Technol.* **36**, 745–760 (2019).
81. de Boyer Montégut, C., Madec, G., Fischer, A. S., Lazar, A. & Iudicone, D. Mixed layer depth over the global ocean: an examination of profile data and a profile-based climatology. *J. Geophys. Res.* **109**, C12003 (2004).
82. Dong, S., Sprintall, J., Gille, S. T. & Talley, L. Southern Ocean mixed-layer depth from Argo float profiles. *J. Geophys. Res.* **113**, C06013 (2008).
83. Marshall, J. & Schott, F. Open-ocean convection: observations, theory, and models. *Rev. Geophys.* **37**, 1–64 (1999).
84. Margirier, F. et al. Characterization of convective plumes associated with oceanic deep convection in the northwestern Mediterranean from high-resolution in situ data collected by gliders. *J. Geophys. Res. Oceans* **122**, 9814–9826 (2017).
85. Haumann, F. A., Gruber, N., Münnich, M., Frenger, I. & Kern, S. Sea-ice transport driving Southern Ocean salinity and its recent trends. *Nature* **537**, 89–92 (2016).
86. Charrassin, J.-B. et al. Southern Ocean frontal structure and sea-ice formation rates revealed by elephant seals. *Proc. Natl Acad. Sci. USA* **105**, 11634–11639 (2008).

87. Bailey, D. A., Rhines, P. B. & Häkkinen, S. Formation and pathways of North Atlantic Deep Water in a coupled ice–ocean model of the Arctic–North Atlantic Oceans. *Clim. Dyn.* **25**, 497–516 (2005).
88. Frajka-Williams, E., Rhines, P. B. & Eriksen, C. C. Horizontal stratification during deep convection in the Labrador Sea. *J. Phys. Oceanogr.* **44**, 220–228 (2014).
89. Pellichero, V., Sallée, J.-B., Schmidtko, S., Roquet, F. & Charrassin, J.-B. The ocean mixed layer under Southern Ocean sea-ice: seasonal cycle and forcing. *J. Geophys. Res. Oceans* **122**, 1608–1633 (2017).
90. Talley, L. D., Pickard, G. L., Emery, W. J. & Swift, J. H. *Descriptive Physical Oceanography: An Introduction* Ch. 7, 187–222 (Elsevier, 2011).
91. Gouretski, V. World Ocean Circulation Experiment – Argo Global Hydrographic Climatology. *Ocean Sci.* **14**, 1127–1146 (2018).
92. Fahrbach, E. et al. Warming of deep and abyssal water masses along the Greenwich meridian on decadal time scales: the Weddell gyre as a heat buffer. *Deep Sea Res. Part II* **58**, 2509–2523 (2011).
93. Ryan, S., Schröder, M., Huhn, O. & Timmermann, R. On the warm inflow at the eastern boundary of the Weddell Gyre. *Deep Sea Res. Part I* **107**, 70–81 (2016).
94. Smedsrud, L. H. Warming of the deep water in the Weddell Sea along the Greenwich meridian: 1977–2001. *Deep Sea Res. Part I* **52**, 241–258 (2005).
95. Fahrbach, E., Hoppema, M., Rohardt, G., Schröder, M. & Wisotzki, A. Causes of deep-water variation: comment on the paper by L.H. Smedsrud “Warming of the deep water in the Weddell Sea along the Greenwich meridian: 1977–2001”. *Deep Sea Res. Part I* **53**, 574–577 (2006).
96. Gordon, A. L. Weddell Deep Water variability. *J. Mar. Res.* **40**, 199–217 (1982).
97. Zanowski, H. & Hallberg, R. Weddell Polynya transport mechanisms in the abyssal ocean. *J. Phys. Oceanogr.* **47**, 2907–2925 (2017).
98. Marshall, G. J. Trends in the Southern Annular Mode from observations and reanalyses. *J. Clim.* **16**, 4134–4143 (2003).
99. Dee, D. P. et al. The ERA-Interim reanalysis: configuration and performance of the data assimilation system. *Q. J. R. Meteorol. Soc.* **137**, 553–597 (2011).
100. Sumata, H. et al. An intercomparison of Arctic ice drift products to deduce uncertainty estimates. *J. Geophys. Res. Oceans* **119**, 4887–4921 (2014).
101. Martinson, D. G. & Wamser, C. Ice drift and momentum exchange in winter Antarctic pack ice. *J. Geophys. Res.* **95**, 1741–1755 (1990).
102. Wang, Z., Turner, J., Sun, B., Li, B. & Liu, C. Cyclone-induced rapid creation of extreme Antarctic sea ice conditions. *Sci. Rep.* **4**, 5317 (2015).
103. Kottmeier, C. & Sellmann, L. Atmospheric and oceanic forcing of Weddell Sea ice motion. *J. Geophys. Res. Oceans* **101**, 20809–20824 (1996).
104. Fairall, C. W., Bradley, E. F., Rogers, D. P., Edson, J. B. & Young, G. S. Bulk parameterization of air-sea fluxes for Tropical Ocean-Global Atmosphere Coupled-Ocean Atmosphere Response Experiment. *J. Geophys. Res. Oceans* **101**, 3747–3764 (1996).
105. Renfrew, I. A., Moore, G. W. K., Guest, P. S. & Bumke, K. A comparison of surface layer and surface turbulent flux observations over the Labrador Sea with ECMWF analyses and NCEP reanalyses. *J. Phys. Oceanogr.* **32**, 383–400 (2002).
106. Holland, P. R. & Kwok, R. Wind-driven trends in Antarctic sea-ice drift. *Nat. Geosci.* **5**, 872–875 (2012).
107. Jullion, L., Jones, S. C., Naveira Garabato, A. C. & Meredith, M. P. Wind-controlled export of Antarctic Bottom Water from the Weddell Sea. *Geophys. Res. Lett.* **37**, L09609 (2010).
108. Meijers, A. J. S. et al. Wind-driven export of Weddell Sea slope water. *J. Geophys. Res. Oceans* **121**, 7530–7546 (2016).
109. Armitage, T. W. K., Kwok, R., Thompson, A. F. & Cunningham, G. Dynamic topography and sea level anomalies of the Southern Ocean: variability and teleconnections. *J. Geophys. Res. Oceans* **123**, 613–630 (2018).
110. Turner, J. et al. The SCAR READER project: toward a high-quality database of mean Antarctic meteorological observations. *J. Clim.* **17**, 2890–2898 (2004).
111. Smith, A., Lott, N. & Vose, R. The Integrated Surface Database: recent developments and partnerships. *Bull. Am. Meteorol. Soc.* **92**, 704–708 (2011).
112. Bracegirdle, T. J. Climatology and recent increase of westerly winds over the Amundsen Sea derived from six reanalyses. *Int. J. Climatol.* **33**, 843–851 (2013).
113. Patoux, J., Yuan, X. & Li, C. Satellite-based midlatitude cyclone statistics over the Southern Ocean: 1. Scatterometer-derived pressure fields and storm tracking. *J. Geophys. Res.* **114**, D04105 (2009).
114. Amante, C. & Eakins, B. W. *ETOPO1 1 Arc-Minute Global Relief Model: Procedures, Data Sources and Analysis*. NOAA Technical Memorandum NESDIS NGDC-24 (National Geophysical Data Center, 2009) <https://doi.org/10.7289/V5C8276M>.

Acknowledgements We thank A. Wong, R. Drucker, J. Plant, T. Maurer and K. Johnson for assistance with float data calibration, and all others involved in float and sensor design, construction, calibration and deployment for their contributions. Data were collected and made freely available by the Southern Ocean Carbon and Climate Observations and Modeling (SOCCOM) Project, which is funded by the US National Science Foundation, Division of Polar Programs (NSF PLR-1425989), supplemented by NASA, and by the International Argo Program and the NOAA programmes that contribute to it. The Argo Program is part of the Global Ocean Observing System. E.C.C. acknowledges funding from the University of Washington (UW) Program on Climate Change, ARCS Foundation, and US Department of Defense through the National Defense Science & Engineering Graduate (NDSEG) Fellowship Program. E.A.W., S.C.R., M.R.M. and L.D.T. acknowledge funding from NSF PLR-1425989, and S.C.R. from NOAA grant NA15OAR4320063. G.W.K.M. acknowledges the support of the Canada Fulbright Foundation, UW Jackson School of International Studies, and the Natural Sciences and Engineering Research Council of Canada. C.E.B. was supported by the Scripps Undergraduate Research Fellowship (SURF) programme.

Author contributions E.C.C., E.A.W. and S.C.R. conceived the study and E.C.C. wrote the initial manuscript. E.C.C. and E.A.W. analysed the hydrographic data and together with G.W.K.M. analysed the sea ice and reanalysis data. S.C.R. led the float design and construction and together with L.D.T. coordinated SOCCOM float deployments. All authors interpreted results and provided input to the final manuscript.

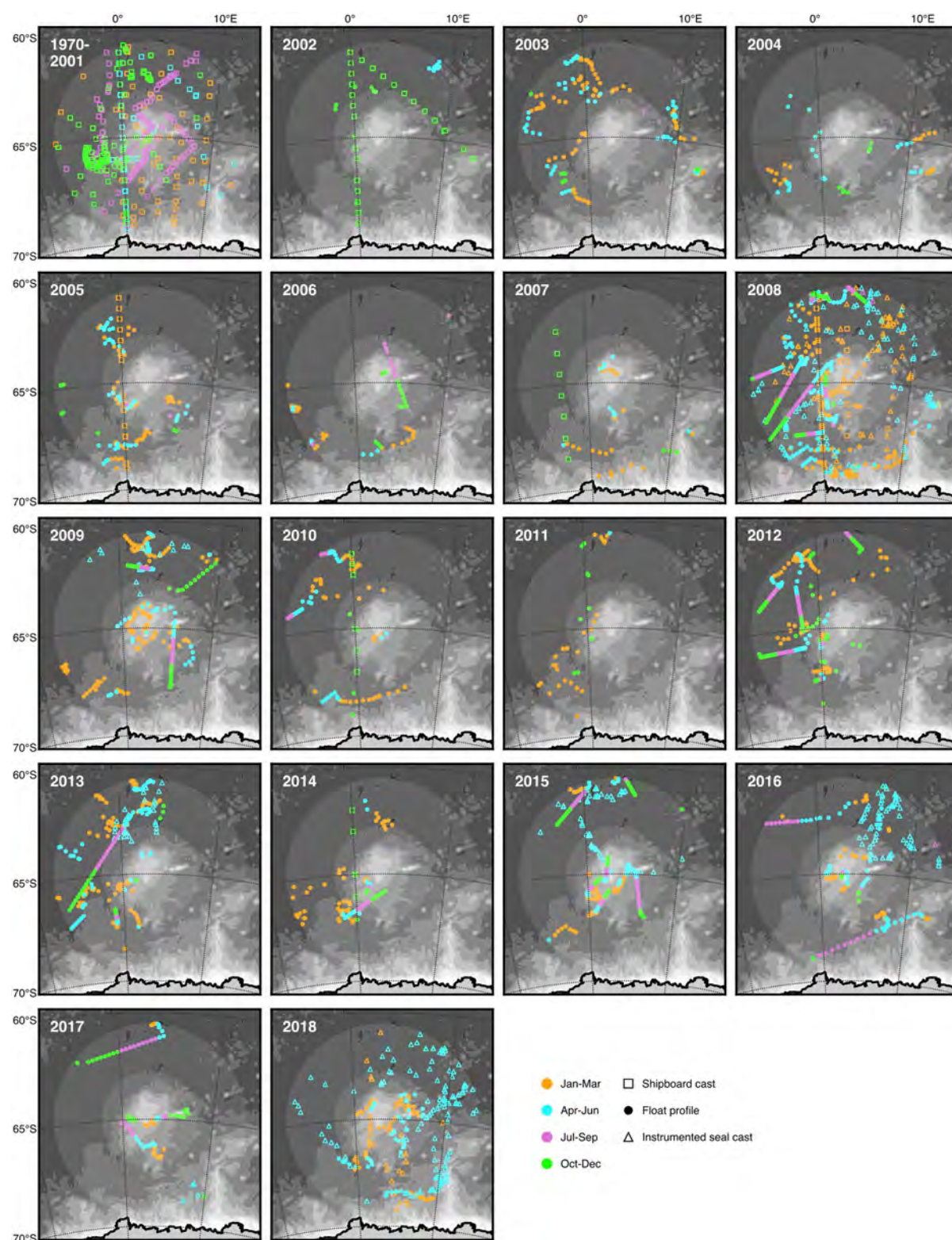
Competing interests The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to E.C.C.

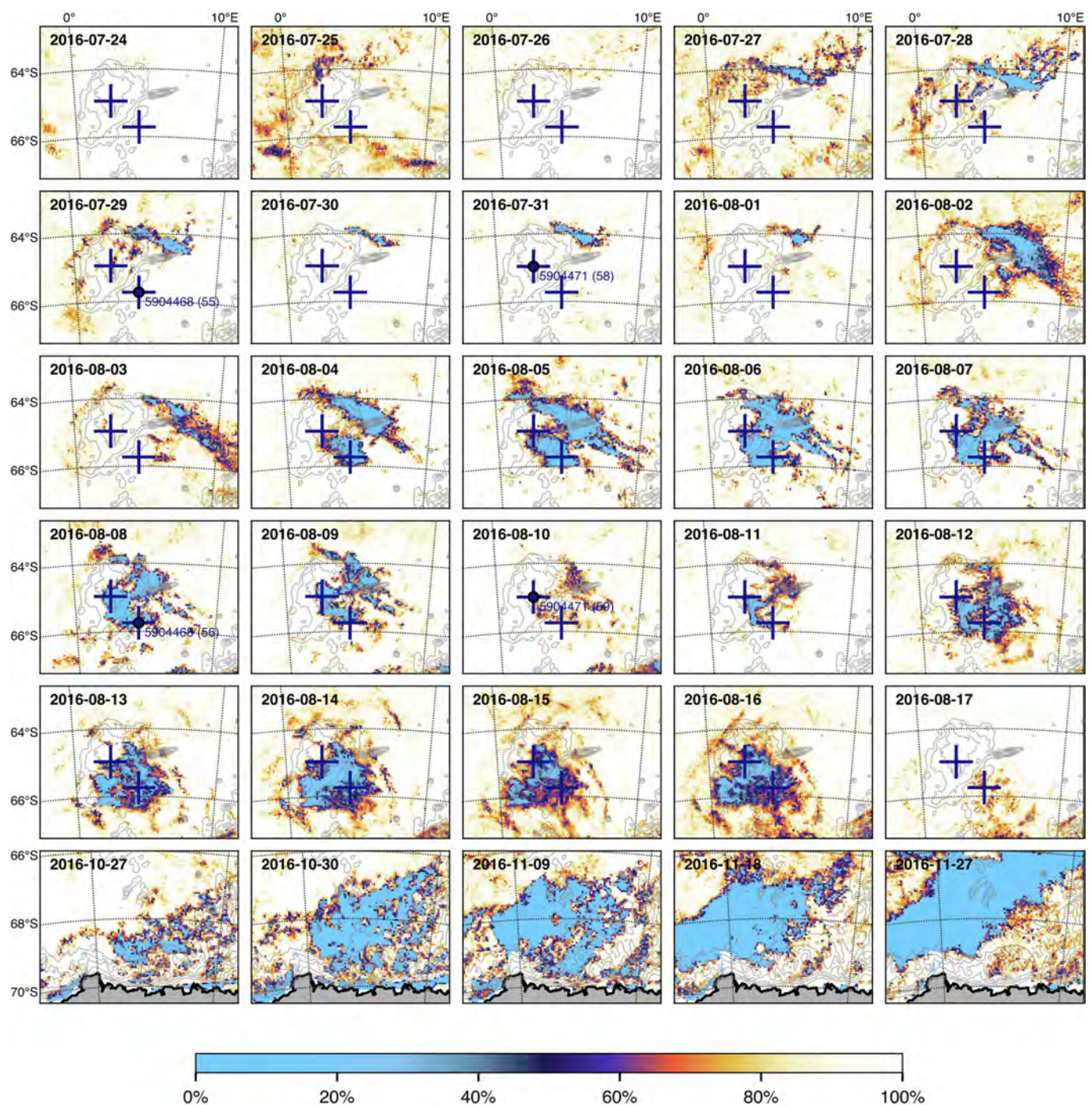
Peer review information *Nature* thanks Céline Heuzé, Lars Henrik Smedsrud and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at <http://www.nature.com/reprints>.



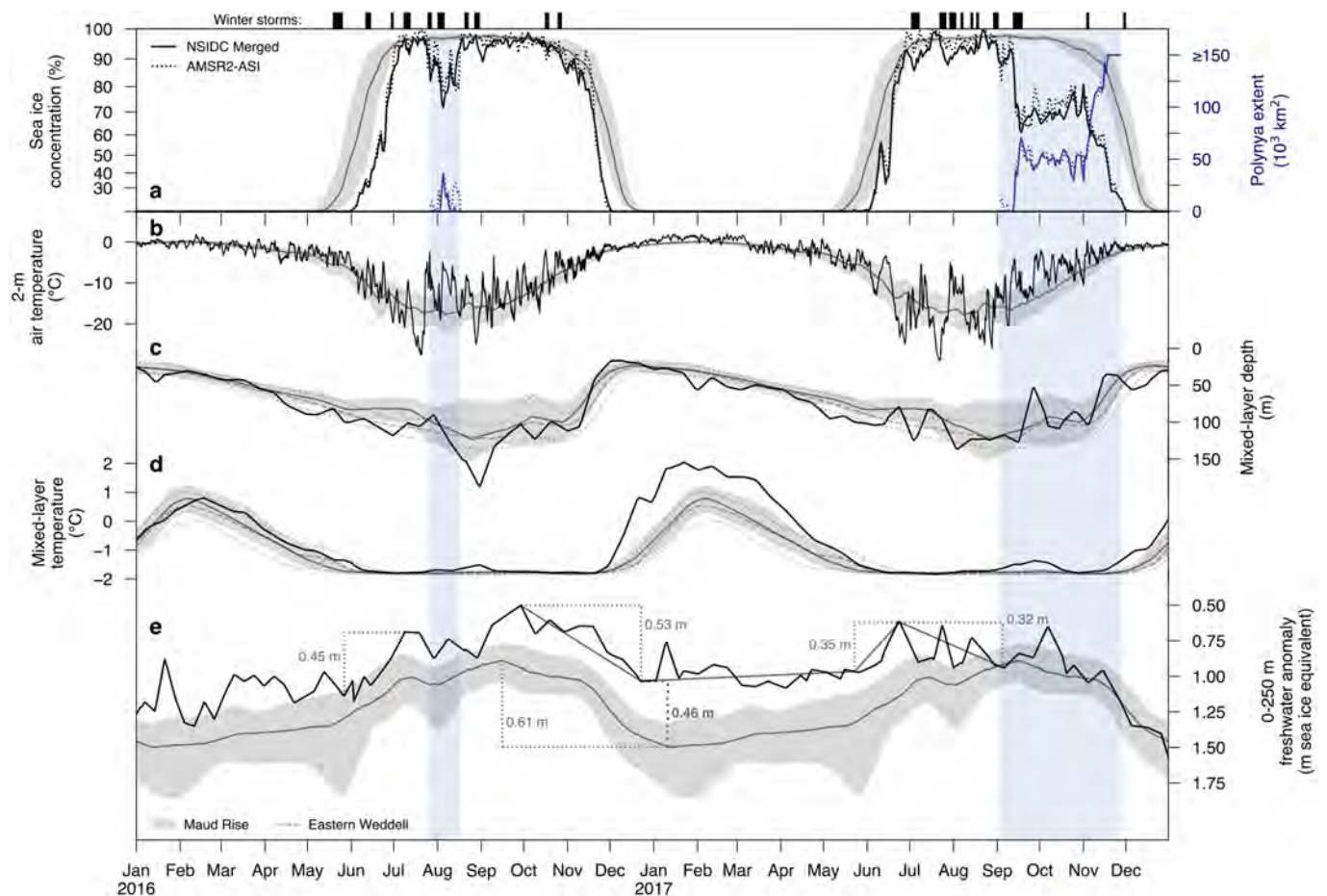
Extended Data Fig. 1 | Locations of observations used to construct hydrographic climatologies for the Maud Rise and eastern Weddell regions. Observations from 1970 to 2001 are shown together (top left); observations from 2002 to 2018 are represented by one panel per year. Included are float profiles from the Argo GDAC (filled circles) as well as shipboard (open squares) and instrumented seal (open triangles) casts

from the World Ocean Database (see Methods section ‘Hydrographic data’). Colours indicate seasons. Bathymetric contours (intervals of 750 m) highlight Maud Rise and the Antarctic continental shelf. Concentric circles represent radii of 250 km and 500 km from Maud Rise, encompassing the Maud Rise and eastern Weddell regions, respectively (see Methods section ‘Regions’).



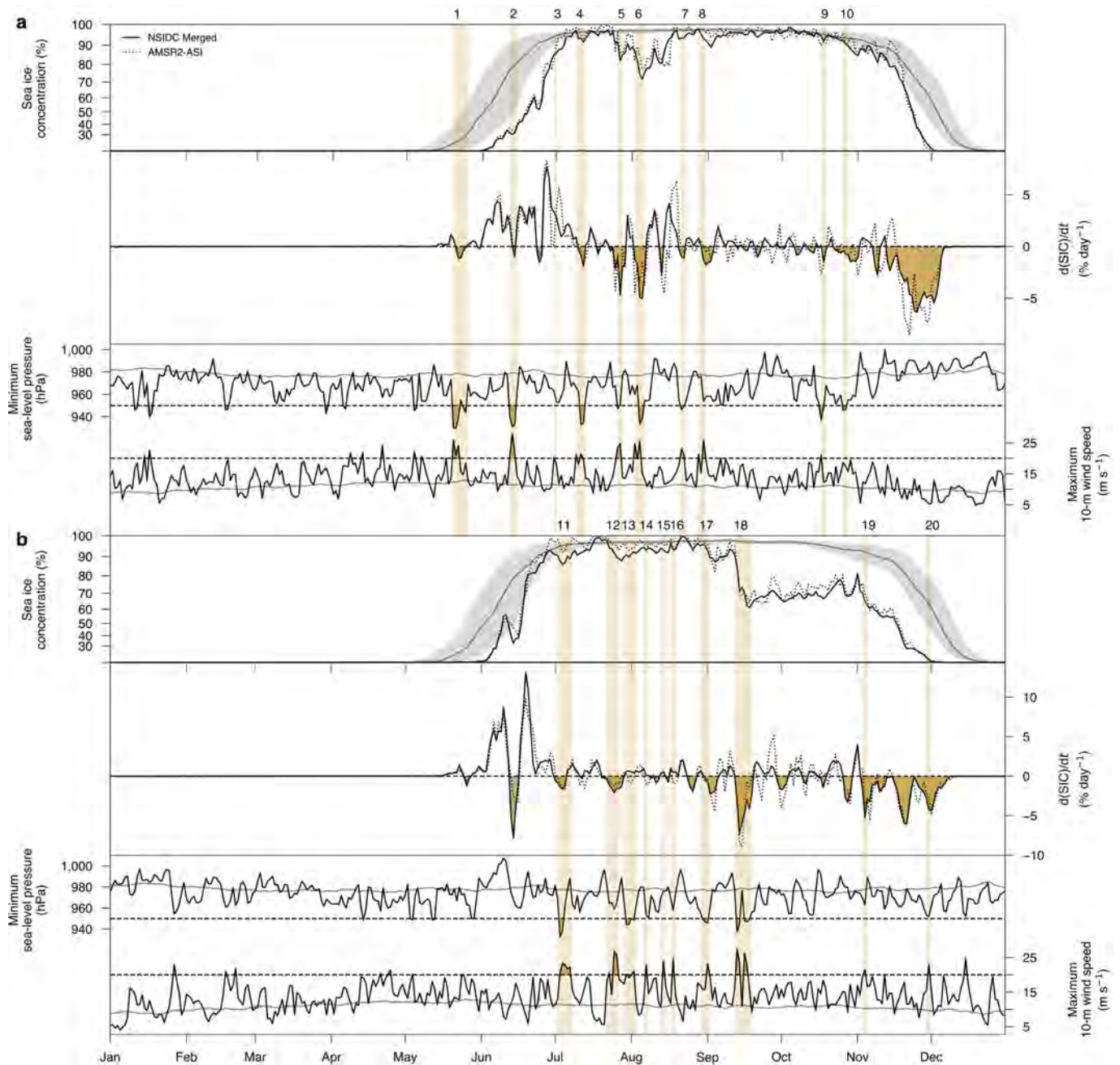
Extended Data Fig. 2 | Sea ice concentration during the 2016 polynya. Daily SIC from AMSR2-ASI around Maud Rise from 24 July to 17 August 2016, encompassing the main polynya event, followed by selected SIC fields from AMSR2-ASI during the late-winter 2016 polynya south of Maud Rise (bottom row; note different map area). Estimated locations of SOCCOM profiling floats 5904471 and 5904468 (see Methods section

'Hydrographic data') are marked in blue; a circle marker and profile number indicate that a hydrographic profile was obtained on that date. Bathymetry shallower than 3,500 m is contoured at intervals of 500 m to highlight Maud Rise (centre of July–August images) as well as Astrid Ridge (bottom right of October–November images), an extension of the Antarctic continental shelf.



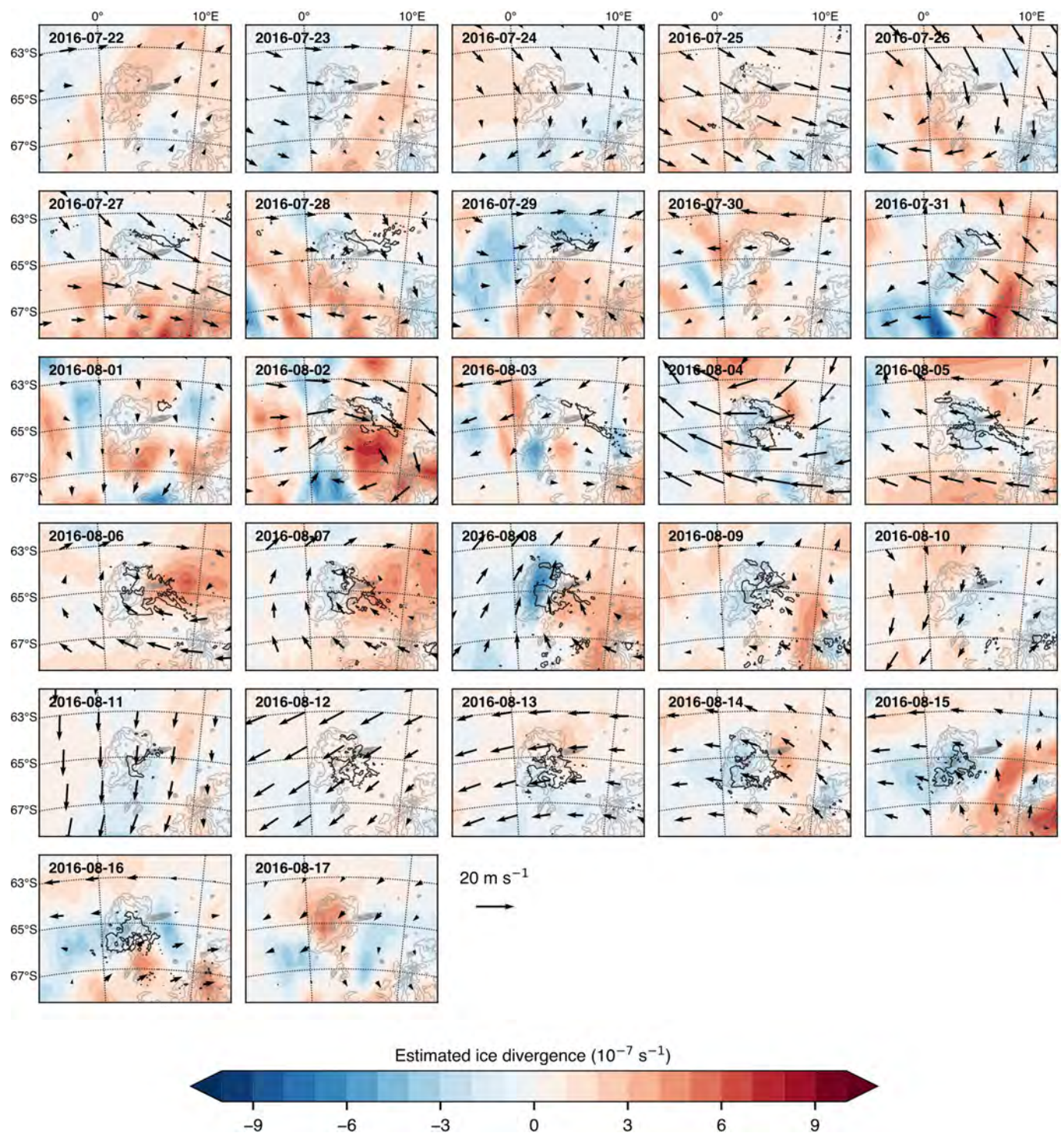
Extended Data Fig. 3 | Evolution of sea ice concentration, air temperature and upper ocean properties at Maud Rise in 2016 and 2017. Marked at the top are intense winter storm events near Maud Rise, as in Fig. 2 (also see Extended Data Fig. 4 and Methods section ‘Storm identification’). **a**, Average daily SIC within the Maud Rise region (63° – 67° S, 0° – 10° E) from NSIDC Merged (solid black line) and AMSR2-ASI (dashed black line) in 2016 and 2017, as in Fig. 2a. SIC climatology from NSIDC Merged (1978–2019) is shown as median (grey line) and 25–75% interquartile range (IQR; grey shading). Note the stretched y axis. Polynya extent is quantified (blue lines) during the 2016 and 2017 events (vertical blue shading). **b**, Six-hourly 2-m air temperature around Maud Rise (within 63° – 67° S, 0° – 10° E) from ERA-I reanalysis (black line). Climatology for 1979–2018 is shown as mean (grey line) and IQR (grey shading). **c**, Composite of average MLD in 2016 and 2017 measured by floats 5903616, 5904468 and 5904471 (black line; see Methods sections

‘Derived oceanographic quantities’ and ‘Composites of float time series’). MLD climatology for the Maud Rise region ($R < 250$ km from 65° S, 3° E) is shown as median (grey line) and IQR (grey shading); climatology for the eastern Weddell region away from Maud Rise ($250 < R < 500$ km) is presented for comparison (light brown dashed and dotted lines for median and IQR, respectively; see Methods section ‘Hydrographic climatologies’). **d**, Composite of average mixed-layer potential temperature (MLT) and MLT climatology presented as in **c**. **e**, Composite of the lowest observed upper-250-m freshwater anomaly (or ‘salt deficit’; see Methods sections ‘Derived oceanographic quantities’ and ‘Composites of float time series’) and freshwater anomaly climatology for the Maud Rise region presented as in **c**. Note the reversed y axis. Key changes quantified, from left to right, are 2016 freeze, climatological melt, 2016 melt, anomaly from climatology in January 2017, 2017 freeze, and change between 2017 freeze and 2017 polynya appearance.



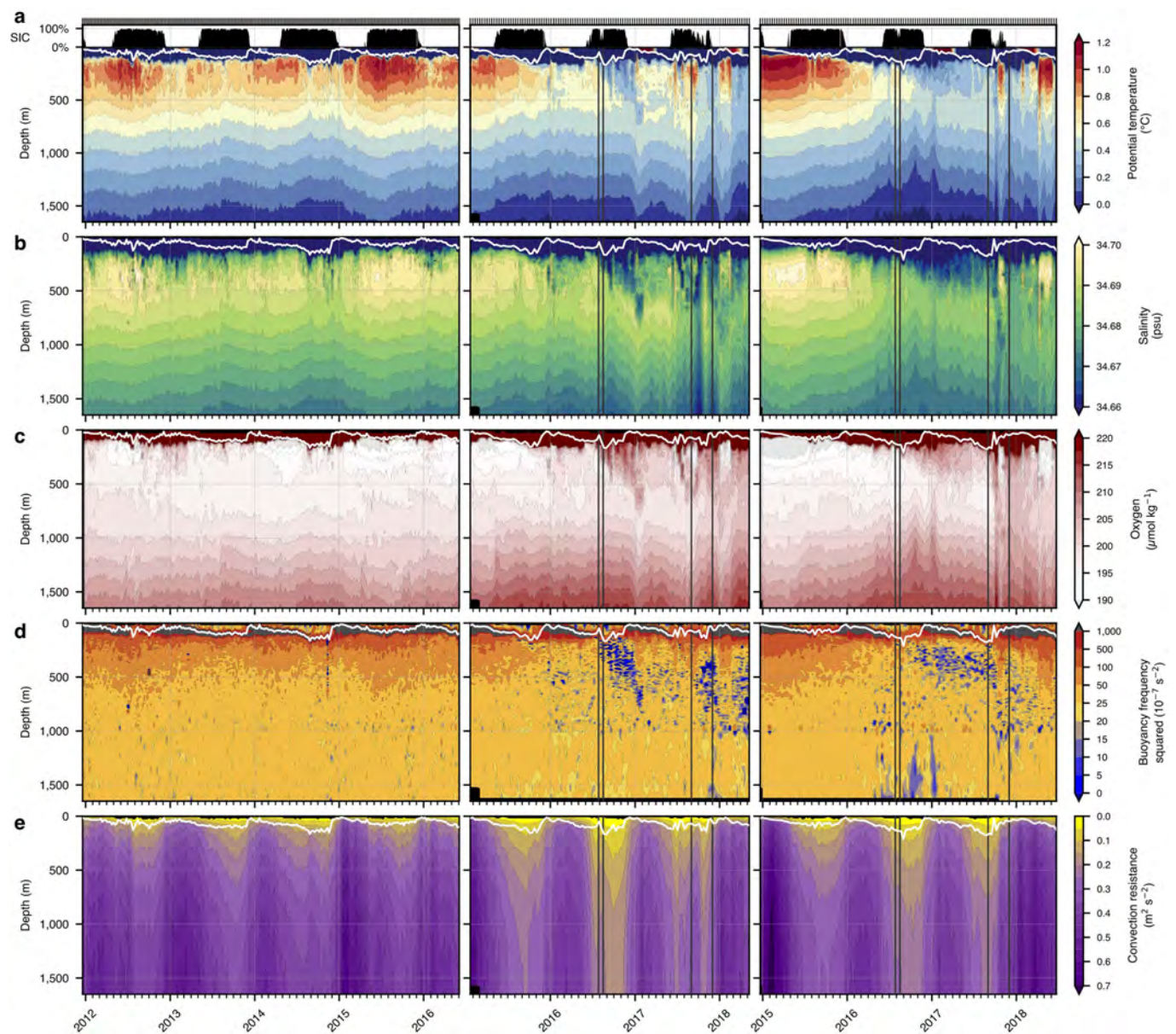
Extended Data Fig. 4 | Correspondence of sea ice loss episodes and major storms near Maud Rise. a, b, Time series are shown for 2016 (a) and 2017 (b). Average daily SIC within the Maud Rise region (63° – 67° S, 0° – 10° E) from NSIDC Merged (solid black line) and AMSR2-ASI (dashed black line) is presented at the top for each year, as in Fig. 2a. SIC climatology from NSIDC Merged (1978–2019) is shown as median (grey line) and 25–75% interquartile range (grey shading). Note the stretched y axis. Daily changes in SIC are presented in the centre for each year, with negative changes in NSIDC Merged highlighted (dark yellow shading); both NSIDC Merged and AMSR2-ASI time series are smoothed using a

3-day right-edge running-mean filter. At the bottom for each year are the minimum daily sea-level pressure and maximum daily 10-m wind speed near Maud Rise (within 63° – 67° S, 0° – 10° E) from ERA-I reanalysis. Mean climatological values of these minimum/maximum metrics are shown (grey lines) to highlight the lack of a pronounced seasonal cycle. The most intense winter polar lows, as shown in Figs. 2, 3, Extended Data Fig. 3, are identified here using pressure and wind speed thresholds (dark yellow shading from dashed lines), and aggregated ‘storm days’ are numbered at the top and marked with vertical yellow bars (see Methods section ‘Storm identification’).



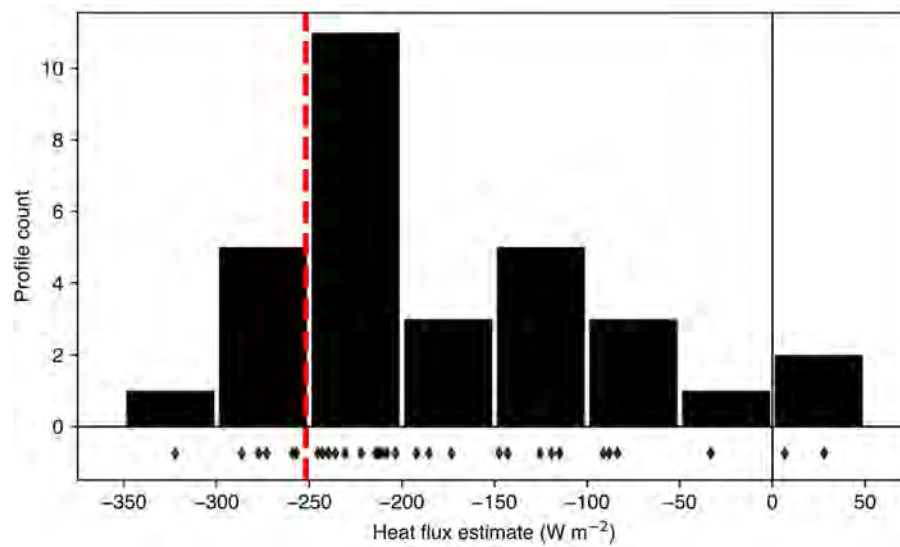
Extended Data Fig. 5 | Winds and wind-induced sea ice divergence during the 2016 polynya. 50% SIC contours (black) from AMSR2-ASI show the daily polynya evolution from 22 July to 17 August 2016. Bathymetry shallower than 3,500 m is contoured at intervals of 500 m (light grey) to highlight Maud Rise (centre). Daily mean 10-m wind

vectors from ERA-I reanalysis, subsampled as every fifth u -wind and every second v -wind vector, are plotted with a 20 m s^{-1} key as reference. Estimated daily mean wind-induced sea ice divergence (see Methods section 'Atmospheric reanalysis') is shaded such that red represents divergence and blue represents convergence.



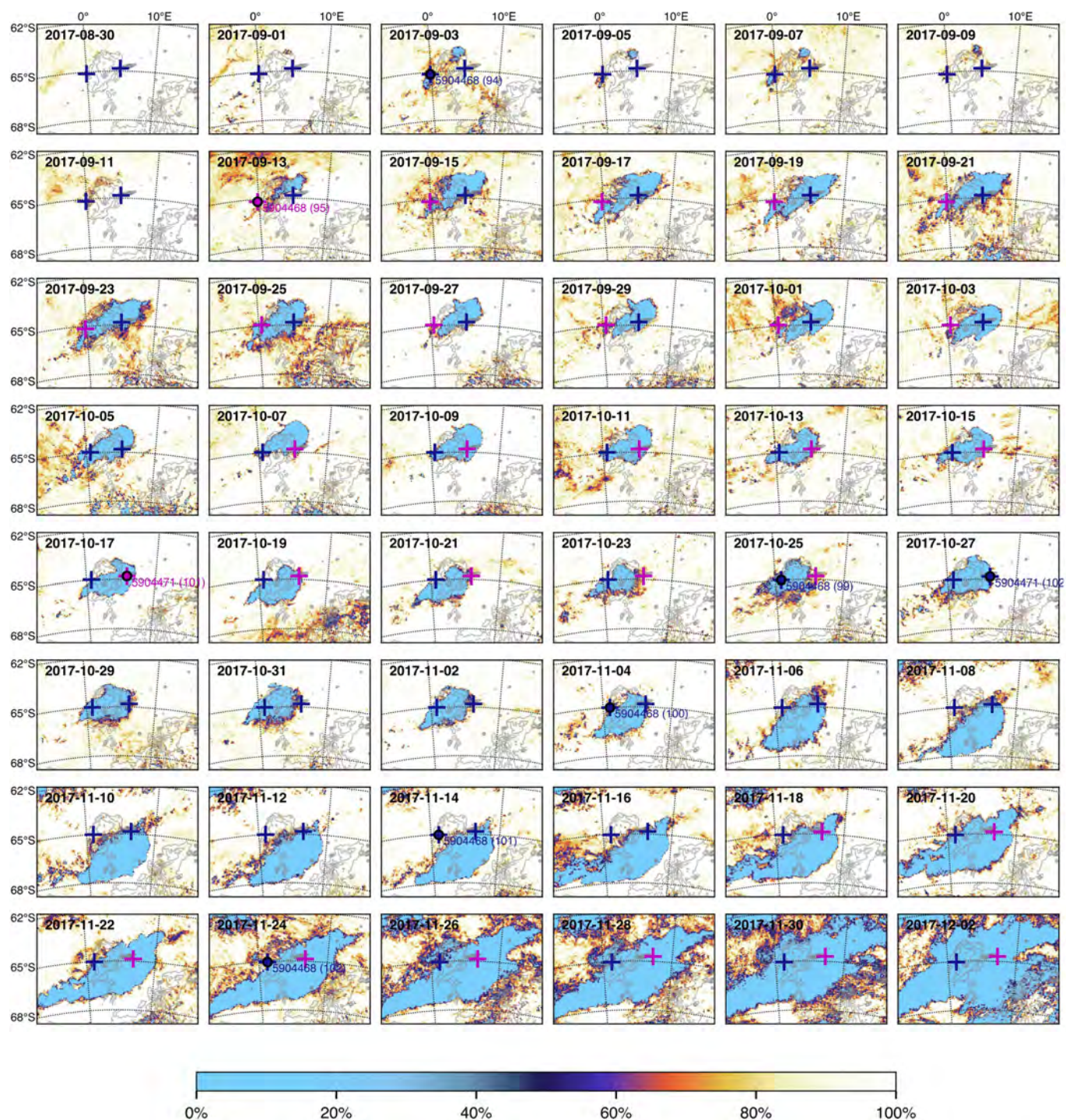
Extended Data Fig. 6 | Full set of profiling float hydrographic observations from Maud Rise from 2011–2018. a–e, Complete depth sections of potential temperature (a), salinity (b), dissolved oxygen (c), buoyancy frequency squared (N^2) (d) and convection resistance (see Methods section ‘Derived oceanographic quantities’) (e) from profiling floats 5903616 (left), 5904468 (centre) and 5904471 (right), as

shown in Fig. 4. Individual profiles are marked at the top (black ticks). Mixed-layer depth is indicated in white. Vertical lines in each panel mark the start and end dates of the 2016 and 2017 polynyas. Along-trajectory SIC, primarily from AMSR2-ASI, is shaded at the top in black (see Methods section ‘Sea ice concentration data’).



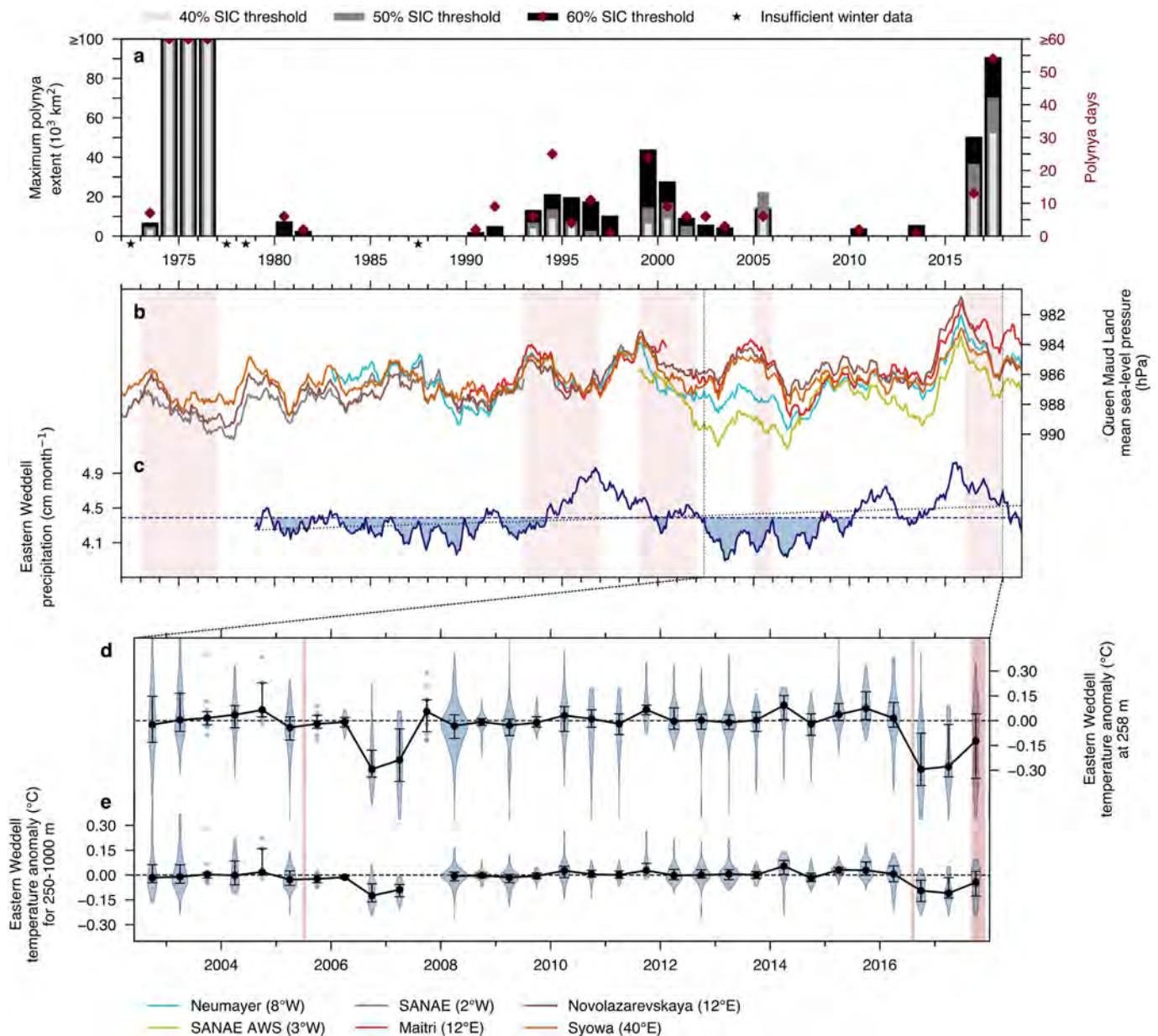
Extended Data Fig. 7 | Heat loss during the 2016 polynya estimated from hydrographic observations. Heat flux estimates ($n = 31$; diamonds at the bottom correspond to histogram above) computed using potential temperature profiles from floats 5904468 and 5904471 following the opening of the 2016 polynya (see Methods section 'Polynya heat flux

estimates'). The dashed red line marks the average open-water ocean-atmosphere turbulent heat flux within the 2016 opening, estimated using a bulk flux algorithm as 252 W m^{-2} (see Methods section 'Atmospheric reanalysis').



Extended Data Fig. 8 | Sea ice concentration during the 2017 polynya. SIC from AMSR2-ASI around Maud Rise is shown every other day from 30 August to 2 December 2017. Estimated locations of SOCCOM profiling floats 5904471 and 5904468 are marked in blue; locations of profiling floats following an ice-free profile with a known position fix are marked

in pink (see Methods section 'Hydrographic data'). A circle marker and profile number indicate that a hydrographic profile was obtained on that date. Bathymetry shallower than 3,500 m is contoured at intervals of 500 m to highlight Maud Rise (centre) as well as Astrid Ridge (bottom right), an extension of the Antarctic continental shelf.



Extended Data Fig. 9 | Additional relationships between past polynyas near Maud Rise, climate forcing, and sub-pycnocline temperatures.

a, Annual maximum polynya extent (bars) and number of polynya days (red diamonds; see Methods section ‘Polynya identification’), as in Fig. 5a. Maximum polynya extent is calculated for three SIC thresholds representing increasingly strict polynya definitions: 60%, 50% and 40%. Polynya days are quantified using the 60% threshold. Stars indicate years with incomplete or absent SIC records. Years with polynya activity at the 50% threshold are shaded vertically in red in **b**, **c**, and likewise in **d**, **e**, except vertical shading delimits the actual major polynya events. **b**, Mean sea-level pressure records from Queen Maud Land meteorological stations, 1972–2018 (see legend and Methods section ‘Meteorological station records’). **c**, Eastern Weddell average precipitation from ERA-I reanalysis between 1979 and 2018, shaded below its mean value to indicate

years with polynya-favourable conditions (that is, lower atmosphere–ocean freshwater flux), consistent with Fig. 5. Time series in **b** and **c** are filtered using a two-year centred running mean to highlight longer-term fluctuations. **d**, Biannually binned eastern Weddell region (within 500 km of Maud Rise) shipboard, float and instrumented seal temperature observations at 258 m from 2002 to 2017, expressed as anomalies from WAGHC gridded hydrographic climatology (see Methods section ‘Sub-pycnocline temperature records’). Error bars denote median and 25–75% IQR. Violin plots summarize the data distribution for $n > 10$, and individual anomalies are shown for $5 \leq n \leq 10$. Periods with $n < 5$ are not plotted. **e**, As in **d**, but showing average temperature anomalies from WAGHC climatology between 250–1,000 m (see Methods section ‘Sub-pycnocline temperature records’). See Extended Data Table 1 for trends and significance for **c–e**.

Extended Data Table 1 | Correlations and trends for climate indices and sub-pycnocline temperature records

	1	2	3	4	5	6	7	Period	Trend (decade ⁻¹)	<i>p</i>
1 Southern Annular Mode (SAM) index [†]	1.00 (0)	-0.71 (0)	-0.82 (0)	-0.38 (-1)	-0.51 (0)	0.30 (-1)	0.36 (-1)	1972-2018	0.25	0.00
2 Weddell Low SLP		1.00 (0)	0.84 (0)	0.48 (0)	0.61 (0)	-0.55 (-1)	-0.38 (0)	1979-2018	-0.24 hPa	0.13
3 Novolazarevskaya SLP [†]			1.00 (0)	0.48 (0)	0.59 (0)	-0.50 (-1)	-0.31 (0)	1972-2018	-0.81 hPa	0.00
4 Maud Rise wind stress curl				1.00 (0)	0.50 (0)	-0.40 (-2)	-0.56 (0)	1979-2018	-0.07·10 ⁻⁷ N m ⁻³	0.16
5 Weddell gyre wind stress curl [†]					1.00 (0)	-0.47 (-1)	-0.67 (0)	1979-2018	-0.05·10 ⁻⁷ N m ⁻³	0.02
6 E. Weddell winter storm days per month* [†]						1.00 (0)	0.33 (3)	1979-2018	0.50	0.01
7 E. Weddell precipitation							1.00 (0)	1979-2018	0.07 cm month ⁻¹	0.15
E. Weddell 258-m temperature anomaly								Jan. 2008 - Jun. 2016	0.07° C	0.04
								Jul. 2002 - Jun. 2016	0.05° C	0.17
E. Weddell 250-1000-m temperature anomaly								Jan. 2008 - Jun. 2016	0.03° C	0.11
								Jul. 2002 - Jun. 2016	0.03° C	0.03

Pearson correlation coefficients (*r*) between climate indices from Fig. 5 and Extended Data Fig. 9 are listed for the lag within ± 3 years at which the absolute value of the correlation is maximal. Lags are cited within parentheses in months; positive values indicate the index on the top axis leads and the index on the vertical axis lags. Except where noted (see footnotes), correlations were calculated after applying a 12-month centred running mean to mitigate seasonality, with a minimum filter window of 6 months. Linear trend estimates and significance (*p*) were computed using a two-sided Wald test. At the top right, trends are shown for climate indices, assessed using the original, unfiltered time series. At the bottom right, trends are shown for both the full period of sub-pycnocline temperature anomaly records before the 2016 polynya as well as the subperiod beginning in 2008, calculated using the biannual median values (see Extended Data Fig. 9d, e). We analyse the subperiod beginning in 2008 separately because the amount of float and elephant seal observations within the eastern Weddell region increased sharply that year (Extended Data Fig. 1).

*Six-month (winter only) running mean applied.

[†]Series detrended (original series found to have significant trend, that is, two-sided $p < 0.05$).

Atypical behaviour and connectivity in *SHANK3*-mutant macaques

Yang Zhou^{1,2,12,13}, Jitendra Sharma^{3,4,5,6,13}, Qiong Ke^{7,8,13}, Rogier Landman^{2,9,13}, Jingli Yuan¹⁰, Hong Chen⁷, David S. Hayden¹¹, John W. Fisher III¹¹, Mingqiang Jiang¹, William Menegas², Tomomi Aida², Ting Yan¹, Ying Zou¹, Dongdong Xu¹, Shivangi Parmar^{2,3}, Julia B. Hyman^{2,3}, Adrian Fanucci-Kiss^{2,3}, Olivia Meisner^{2,3}, Dongqing Wang^{2,3}, Yan Huang¹⁰, Yaqing Li¹⁰, Yanyang Bai¹, Wenjing Ji¹, Xinqiang Lai⁷, Weiqiang Li^{7,8}, Lihua Huang⁷, Zhonghua Lu¹, Liping Wang¹, Sheeba A. Anteraper^{2,3}, Mriganka Sur^{3,4,5}, Huihui Zhou^{1*}, Andy Peng Xiang^{7,8*}, Robert Desimone^{2,3}, Guoping Feng^{2,3,9*} & Shihua Yang^{10*}

Mutation or disruption of the SH3 and ankyrin repeat domains 3 (*SHANK3*) gene represents a highly penetrant, monogenic risk factor for autism spectrum disorder, and is a cause of Phelan–McDermid syndrome. Recent advances in gene editing have enabled the creation of genetically engineered non-human-primate models, which might better approximate the behavioural and neural phenotypes of autism spectrum disorder than do rodent models, and may lead to more effective treatments. Here we report CRISPR–Cas9-mediated generation of germline-transmissible mutations of *SHANK3* in cynomolgus macaques (*Macaca fascicularis*) and their F1 offspring. Genotyping of somatic cells as well as brain biopsies confirmed mutations in the *SHANK3* gene and reduced levels of *SHANK3* protein in these macaques. Analysis of data from functional magnetic resonance imaging revealed altered local and global connectivity patterns that were indicative of circuit abnormalities. The founder mutants exhibited sleep disturbances, motor deficits and increased repetitive behaviours, as well as social and learning impairments. Together, these results parallel some aspects of the dysfunctions in the *SHANK3* gene and circuits, as well as the behavioural phenotypes, that characterize autism spectrum disorder and Phelan–McDermid syndrome.

SHANK3 encodes major scaffolding proteins at excitatory synapses, coordinates the recruitment of signalling molecules and creates scaffolds for appropriate alignment of glutamatergic neurotransmitter receptors, which promotes the development and maturation of excitatory synapses^{1,2}. Mutation of *SHANK3* accounts for about 1% of idiopathic forms of autism spectrum disorder, and disruption of *SHANK3* is a major cause of neurodevelopmental deficits in Phelan–McDermid syndrome^{3–6}. Patients with a *SHANK3* gene mutation often exhibit a variety of comorbid traits, which include global developmental delay, severe sleep disturbances, lack of speech or severe language delay, and characteristic features of autism spectrum disorder (such as social impairments and stereotypies)^{7–9}. Previous studies in flies, fish and rodents have uncovered impaired synaptic function and several behavioural abnormalities due to loss of *SHANK3*^{2,10,11}. For example, common abnormalities in *Shank3*-mutant mice include self-injury, repetitive grooming, reduced interaction with conspecifics, motor difficulties and increased levels of anxiety. However, it is increasingly apparent that the validity of *Shank3*-mutant rodent models for human patients is limited, in part owing to the fact that the aberrant behavioural phenotypes in mice are found almost exclusively in homozygous mutants and are barely detectable in heterozygous mutants^{2,12}. In addition, social interactions between humans involve integrated cognition and comprehension, which are more closely associated in primates than in rodents^{13–16}. There is therefore an urgent need to develop primate

models of autism spectrum disorder to facilitate neurobiological studies and the development of therapies^{12,13,17}.

Cynomolgus monkeys (*M. fascicularis*) possess a high level of cognitive ability and complex social behaviour, and are closer to humans in terms of their brain structure and function than are rodents^{12,13,16,17}. There has been great interest in using macaques as a non-human-primate model for studying brain disorders^{17–20}. Recent advances in CRISPR–Cas9-mediated gene-editing technology have resulted in an increasingly efficient and reliable method for targeted gene disruption, which is highly suited to the creation of non-human-primate models of autism spectrum disorder^{21,22}. Previous attempts to create a transgenic *SHANK3* macaque model have been hindered by the early death of the mutant founders²³, or by the fact that only a single mutant survived²⁴—group comparisons are not possible with only a single mutant. A viable primate model of *SHANK3* mutation should include validation of the loss of protein isoforms, and functional and behavioural deficits should be present at the group level. Furthermore, germline transmission of CRISPR–Cas9-edited macaque genomes has yet to be fully demonstrated. Here we report the creation of a *SHANK3*-mutant macaque model with a specific target locus that—in human studies—has been linked to Phelan–McDermid syndrome and autism spectrum disorder. We also demonstrate the resultant functional and behavioural abnormalities using assays that can be adapted for testing with autism spectrum disorder and Phelan–McDermid syndrome in human patients. Furthermore, we successfully obtained an

¹Brain Cognition and Brain Disease Institute, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China. ²McGovern Institute for Brain Research, Massachusetts Institute of Technology, Cambridge, MA, USA. ³Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA. ⁴Picower Institute for Learning and Memory, Massachusetts Institute of Technology, Cambridge, MA, USA. ⁵Simons Center for the Social Brain, Massachusetts Institute of Technology, Cambridge, MA, USA. ⁶Athinoula A. Martinos Center for Biomedical Imaging, Department of Radiology, Massachusetts General Hospital, Charlestown, MA, USA. ⁷Center for Stem Cell Biology and Tissue Engineering, Key Laboratory for Stem Cells and Tissue Engineering, Sun Yat-Sen University, Guangzhou, China. ⁸Guangzhou Regenerative Medicine and Health Guangdong Laboratory, Guangzhou, China. ⁹Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA, USA. ¹⁰College of Veterinary Medicine, South China Agricultural University, Guangzhou, China. ¹¹Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA, USA. ¹²Present address: Montreal Neurological Institute & Hospital, Department of Neurology and Neurosurgery, McGill University, Montreal, Quebec, Canada. ¹³These authors contributed equally: Yang Zhou, Jitendra Sharma, Qiong Ke, Rogier Landman. *e-mail: hh.zhou@siat.ac.cn; xiangp@mail.sysu.edu.cn; fengg@mit.edu; yangsh@scau.edu.cn

F1 generation that has a homogenous pattern of *SHANK3* mutation, using sperm from CRISPR-edited founders.

Germline-transmissible *SHANK3* mutation in macaques

We applied a CRISPR–Cas9 strategy that targets exon 21 of the macaque *SHANK3* gene (Extended Data Fig. 2a). Exon 21 is the largest coding region of *SHANK3*, with numerous rare variants and point mutations in individuals with autism spectrum disorder^{2–5,7}. We created indels in exon 21 analogous to the human autism spectrum disorder-linked InsG3680 mutation²⁵ that were previously generated and analysed in mice^{26,27}. *Streptococcus pyogenes* Cas9 and two guide (g)RNAs were introduced to generate the mutation, and the successful creation of insertions or deletions (indels) were verified (Extended Data Fig. 2b, c). We transferred injected embryos into recipient females, and obtained five live newborns (labelled M1 to M5) that carried *SHANK3* mutations (Extended Data Fig. 2d). Four out of the five mutants (M1, M2, M3 and M5) are male, and one mutant (M4) is female. Genotyping of *SHANK3* from the newborns revealed several types of indel mutations in the *SHANK3* gene (Fig. 1a, Extended Data Fig. 3). M2 and M5 did not show a wild-type allele in genomic DNA; they are homozygous and compound-heterozygous, respectively (Fig. 1a, Extended Data Fig. 3). M1, M3 and M4 carried wild-type alleles in about 50% of the sequenced clones and thus are heterozygous, with (M1 and M3) or without (M4) genetic mosaicism (more than one type of indel). The control group consisted of age- and sex-matched macaques from the same colony. The general health measurements of the *SHANK3*-mutant macaques (such as body weight) were not different from controls (Extended Data Fig. 2e). Body condition scores of control and mutant groups were within the normal range for this species.

Sequencing results in exon 21 of the *SHANK3* gene showed multiple genotypes, which suggests that the CRISPR–Cas9-mediated cleavage of *SHANK3* resulted in a mosaicism of indels, as has previously been reported^{21,22}. All of the indels from *SHANK3* mutants caused frameshift and loss-of-function mutations, except for a 12-bp in-frame deletion in mutant M4 and a 96-bp deletion in mutant M3; these led to a loss of 4 and 32 amino acids, respectively, within the proline-rich domain of *SHANK3*, which might be critical for the stability of *SHANK3* protein isoforms and their interaction with scaffolds or receptors^{1,2,7}. To determine how mutations of *SHANK3* affected the protein products of this gene, we biopsied tissue from primary visual cortex and performed western blot analysis. The biopsies were performed after completion of all behavioural tests and magnetic resonance imaging (MRI) scanning. We confirmed a decreased level of the isoforms of the *SHANK3* protein in all five mutant monkeys (Fig. 1b, c, Extended Data Figs. 1 and 4) and thus validated the successful targeted disruption of exon 21, which has previously been linked to synaptic and circuit dysfunction^{2,11,26,27}.

To test for off-target modifications, we amplified and sequenced the top 20 genomic loci that were predicted by the Cas-OFFinder algorithm²⁸. We analysed the sequencing results of these 20 loci from all 5 *SHANK3* mutants and confirmed their wild-type identities (Supplementary Table 1). Our data support the high fidelity of gene editing with CRISPR–Cas9, which aligns with previous surveys of CRISPR-edited model organisms²².

To assess the germline transmission of the mutation, we collected and analysed the DNA of sperm from M2 (homozygous) and M3 (heterozygous and mosaic). We detected *SHANK3* mutations in the DNA of sperm cells of both individuals, and the patterns of mutation were similar to their respective somatic cells (Fig. 1d, Extended Data Figs. 1, 2f). We then performed intracytoplasmic sperm injection into wild-type oocytes using semen from monkey M2 (homozygous), and detected a 40-bp deletion mutation in all fertilized embryos (Fig. 1e, Extended Data Fig. 1). We successfully obtained live births of F1-generation mutant macaques after transferring the fertilized embryos. Genotyping of F1-generation mutant monkeys from the first cohort (labelled F1-1, F1-2 and F1-3) revealed a similar ratio of wild-type to mutant allele (labelled 'del 40 bp') from all three monkeys (Fig. 1f, Extended

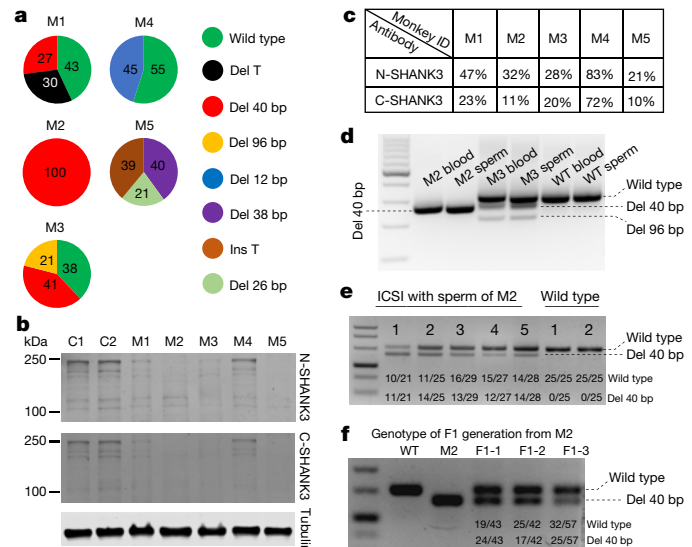


Fig. 1 | Generation and germline transmission of macaques with *SHANK3* mutations. **a**, Pie charts of genotype (indels) from cultured skin fibroblasts derived from each mutant monkey. **b**, Representative western blots using brain lysates prepared from a V1 biopsy of two wild-type macaques (labelled C1 (male) and C2 (female)) and all five mutants (M1–M5), probed with N-terminal and C-terminal antibodies, and α -tubulin as loading control on the same gel. **c**, Relative levels of *SHANK3* proteins calculated by averaging five technical repeats with the same V1 biopsy sample, and normalizing to α -tubulin loading controls. **d**, Genotyping PCR results of *SHANK3* from DNA from blood and sperm samples collected from M2, M3 and a wild-type (WT) control. **e**, Top, genotyping PCR results of *SHANK3* from individual cultured embryos after intracytoplasmic injection of single sperm (ICSI) from M2 and a wild-type control. Bottom, numbers of reads for wild-type and a 40-bp deletion are presented by Sanger sequencing of each bacterial colony after cloning of PCR products into sequencing vector. **f**, Top, genotyping PCR results of three live-birth, F1-generation monkeys after ICSI of sperm from M2. Bottom, numbers of reads for wild-type and a 40-bp deletion for each monkey (F1-1, F1-2 and F1-3) by Sanger sequencing of cloned PCR products.

Data Fig. 1), which indicates that these are heterozygous *SHANK3*-mutant monkeys without genetic mosaicism. Taken together, these results show that the simultaneous delivery of two gRNAs that target both strands of exon 21 enables the efficient mutagenesis of *SHANK3* and the germline transmission of the mutant allele, in cynomolgus monkeys.

Altered sleep, home cage activity and muscle tone

Patients with Phelan–McDermid syndrome and other individuals with *SHANK3* mutations often exhibit a variety of traits, which include motor impairment, severe sleep disturbances, lack of speech or severe language delay, abnormal sensory processing, intellectual disability, muscular hypotonia and seizure, as well as other symptomologies^{2–8}. Most patients with Phelan–McDermid syndrome also exhibit stereotypies, and impairments in social interaction^{4,8}. We used a panel of behavioural tests to examine phenotypes associated with *SHANK3* mutation in macaques. All observations, scoring and data analysis for these tests were carried out by researchers who were blinded to experimental design, goals and genotypes.

To assess sleep disturbance in the mutant monkeys, we habituated them to wearing an actigraphy device. The activity data revealed that overall activity levels were substantially reduced in mutant monkeys compared to controls (Fig. 2a–g). Mutant monkeys displayed a longer latency to sleep than controls, and an increased frequency of waking (as indicated by a fragmentation index) (Fig. 2h, i). These results showed a reduction in overall sleep efficiency in *SHANK3*-mutant monkeys (Fig. 2j, Extended Data Fig. 5).

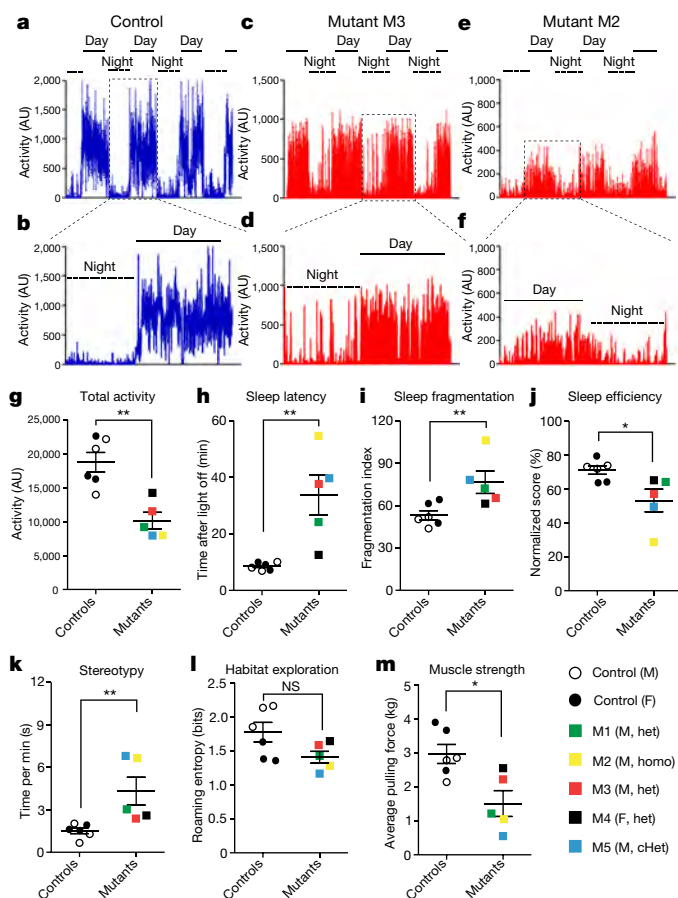


Fig. 2 | Sleep disturbances and altered home-cage activity in *SHANK3*-mutant macaques. **a, c, e,** Representative traces of overall activity recorded by a motion watch across multiple days from a control macaque and two mutants. AU, arbitrary units. **b, d, f,** Enlarged traces of overall activity recorded by motion watch over the course of 24 h from a control macaque, M3 and M2. **g,** Reduced overall activity in *SHANK3* mutants. **h,** Increased latency to fall asleep in *SHANK3* mutants. **i,** Increased fragmentation of sleep in *SHANK3* mutants during the night. **j,** Reduced efficiency of sleep in *SHANK3* mutants during the night. **k,** Increased time spent on stereotypic behaviours, presented as seconds per minute, in *SHANK3* mutants. **l,** *SHANK3* mutants show a trend of reduced area exploration in their home cages, as defined by roaming entropy. **m,** *SHANK3* mutants display reduced muscle strength. In **g–m**, $n = 6$ macaques for control group; $n = 5$ macaques for the *SHANK3*-mutant group; P value = 0.0087 (**g**), 0.0043 for (**h**), 0.0087 (**i**), 0.0303 (**j**), 0.0043 (**k**), 0.08 (**l**) and 0.03 (**m**). Data are presented as mean \pm s.e.m. Two-tailed Mann–Whitney U -test. Coloured squares indicate individual monkeys with *SHANK3* mutation. M, male; F, female; het, heterozygous; homo, homozygous; cHet, compound heterozygous.

We recorded daily home-cage videos and scored behavioural variables using Noldus Observer software (see representative ethograms in Supplementary Table 2). The mutant macaques showed a substantial increase in stereotyped or repetitive behaviours compared to controls, as indicated by increased back-flipping, finger licking and biting of cage bars (Fig. 2k). In contrast to the increased amount of self-injurious grooming that is consistently observed in *Shank3*-mutant mice, the stereotypy in *SHANK3*-mutant monkeys was diverse. For instance, repetitive flipping was observed in mutant M3, whereas M2 and M5 displayed a pronounced licking of fingers and cage bars. The mutant macaques also displayed a trend for reduced level of exploration of cage subdivisions, as demonstrated by roaming entropy (Fig. 2l).

To test for hypotonia, we habituated the monkeys to pull a digital scale and recorded their pulling force. The average pulling force was reduced in the mutant group (Fig. 2m).

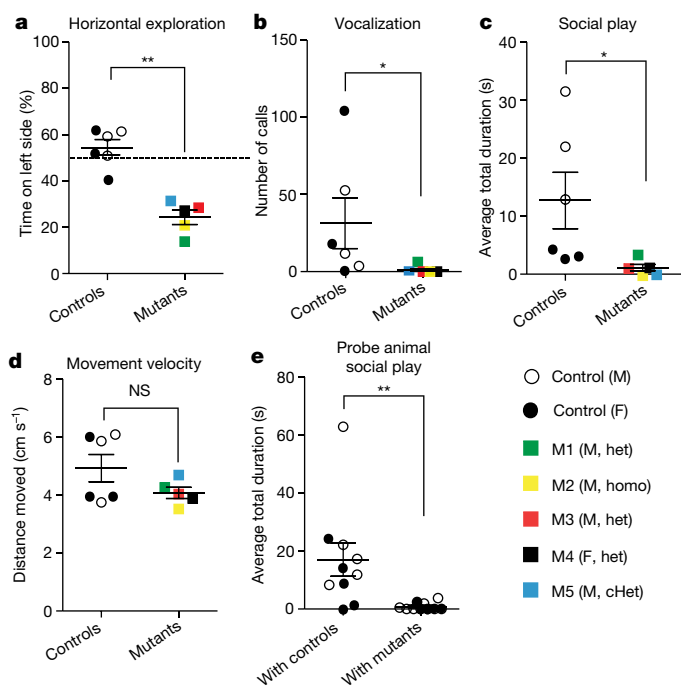


Fig. 3 | Impaired social interaction and reduced vocalization in *SHANK3* mutants. **a,** Proportion of time spent on the left versus right side of the cage shows that mutants tended to stay on one side, whereas control monkeys spent a roughly equal amount of time on each side. **b,** *SHANK3*-mutant monkeys made fewer vocalizations during the habituation period than did controls. **c,** Average total duration of aggregate social behaviours (which included chasing, following, fleeing, circling and play) was lower in mutants than in controls. **d,** No difference in movement velocity during the social test between mutants and controls. **e,** Average total duration of aggregate social behaviours was reduced in probe monkeys (wild type) when paired with *SHANK3* mutants, compared to probe monkeys paired with controls. $n = 6$ macaques for wild-type control group; $n = 5$ macaques for *SHANK3*-mutant group; in **e**, $n = 10$ wild-type macaques (five males, five females and age-matched to experimental animals) for probe monkeys. $P = 0.0043$ (**a**), 0.0444 (**b**), 0.0173 (**c**) and 0.0023 (**e**). NS, not significant. Data are presented as mean \pm s.e.m. Two-tailed Mann–Whitney U -test. Coloured squares indicate individual monkeys with *SHANK3* mutation.

Impaired social interaction

We designed a paired social interaction assay between juvenile monkeys. The monkeys were habituated in the test cage before the social test began (Extended Data Fig. 6a). During habituation, the mutant macaques showed a reduced amount of exploration in the horizontal plane (Fig. 3a). Vocalization during the 20 minutes of habituation was reduced in the mutants (Fig. 3b). In the social interaction test, the scoring of detailed activities for the first five minutes revealed that mutants spent less time on aggregate social behaviours, which included chasing, following, circling, fleeing and play (Fig. 3c, Extended Data Fig. 6b–f). The mean velocity values during the social interactions were similar between the mutant and control monkeys (Fig. 3d), which suggests that physical limitations were unlikely to be major contributing factors to the reduced levels of social interaction. Furthermore, there was no difference in time spent on other categories of behaviours, such as attacking, anogenital inspection, rump presentation, mounting, and receiving and giving grooming during the social test (Extended Data Fig. 6g–l). The aggregate social behaviour values decreased in the subsequent five minutes for the control group, and approached that of the mutant monkeys (Extended Data Fig. 6m). Analysing the behaviours of the wild-type probe monkeys (that is, the monkeys with which the mutant and control groups interacted) during the same time period revealed notable differences between pairings with wild-type controls and the *SHANK3* mutants. In aggregate measures, the probe monkeys

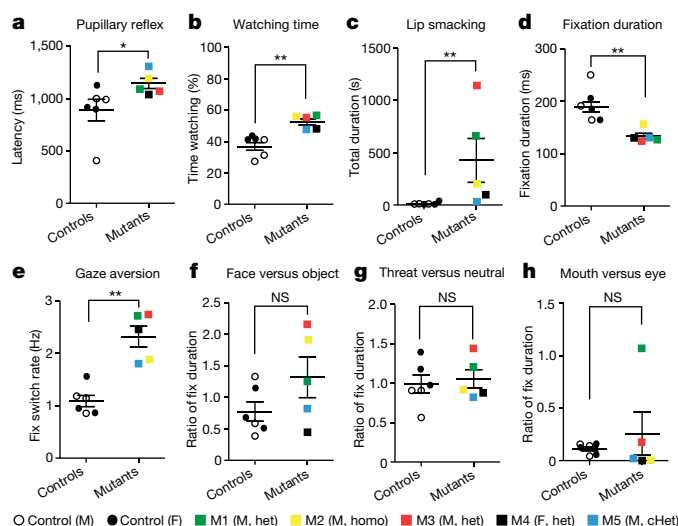


Fig. 4 | Eye-tracking properties in *SHANK3*-mutant macaques.

a, Latency of pupillary reflex upon stimuli illuminance is increased in *SHANK3* mutants. **b**, Percentage of time watching the screen is increased in *SHANK3* mutants. **c**, *SHANK3* mutants display increased lip smacking when watching the video presentation of close-up and whole-body videos of monkeys. **d**, The duration of each fixation during the presentation of images was reduced in *SHANK3* mutants. **e**, Increased fixation-switch rate in *SHANK3* mutants. **f–h**, No difference in the ratio of fixation duration in face versus object (**f**), threat versus neutral faces (**g**) or mouth versus eye region (**h**) from macaques tested with image stimuli. NS, not significant. $n = 6$ macaques for control group; $n = 5$ macaques for mutant group. $P = 0.0303$ (**a**), 0.0043 (**b**), 0.0043 (**c**), 0.008 (**d**) and 0.0043 (**e**). Data are presented as mean \pm s.e.m. Two-tailed Mann–Whitney U -test. Coloured squares indicate individual monkeys with *SHANK3* mutation.

spent less time playing with the *SHANK3* mutants than they did playing with the controls (Fig. 3e, Extended Data Fig. 7a–k). The differential effects on probe monkeys are reminiscent of the effects of oxytocin and vasopressin application on looking behaviour in macaques²⁹. These results suggest the *SHANK3*-mutant monkeys display reduced levels of exploration, social interaction and vocalization, which seem to parallel some aspects of the phenotypes that are found in humans with Phelan–McDermid syndrome or autism spectrum disorder.

Altered gaze properties

We designed a video-based eye-tracking assay to investigate differences in gaze between control and mutant monkeys^{30,31}. The mutants showed an increased latency of pupillary reflex to the onset of luminance at the start of the video stimulus (Fig. 4a), which is consistent with previous reports of delayed pupil reflex in humans with autism spectrum disorder³². The mutant monkeys spent more time watching both social and non-social videos than did controls (Fig. 4b). However, during the presentation of close-up and whole-body videos, mutant monkeys displayed frequent lip smacking and teeth chattering, which could be an indication of increased levels of anxiety or fear in response to the appearance of a conspecific^{33,34} (Fig. 4c). We also assessed eye movements while the monkeys viewed pairwise still images of macaque faces and objects, or faces with threatening and neutral expression^{35,36}. Across all images, the mean dwell time per fixation was reduced in the mutant group (Fig. 4d). There was an increased rate of switching fixations between the two images (Fig. 4e), but not in the ratio of total fixation time (Fig. 4f–h). It has also previously been shown that human children who later develop autism have shorter durations of fixation³⁷.

To evaluate the cognitive performance of mutant monkeys, we trained them to perform a visual discrimination task, a reversal task and a Hamilton search task¹⁸. Despite extensive training, mutants M2 and M5 failed to participate in the test—possibly owing to impaired motor coordination or cognitive deficits. In black–white discrimination and reversal tasks, the controls and remaining mutants (M1, M3 and

M4) performed similarly (Extended Data Fig. 8). In a Hamilton search task with three phases (Extended Data Fig. 9a), mutant and control monkeys were similar in first phase. However, M3 did not improve in the second phase (Extended Data Fig. 9b) and in the third phase, M1 and M4 exhibited learning impairment (Extended Data Fig. 9c). A χ^2 test of the proportion of monkeys that met a criterion of being correct 75% of the time by the final day of testing revealed a difference between the six controls and three mutants (Extended Data Fig. 9d). These results suggest a possible impairment in the mutants in terms of their switching strategies across the phases of Hamilton search; however, the small number of mutant monkeys that could perform the task prevent us from reaching a firm conclusion regarding the nature of their impairment.

Altered brain connectivity

MRI studies of individuals with autism spectrum disorder suggest that considerable structural and functional changes of the brain are associated with this disorder^{38,39}. Structural MRI analysis in *SHANK3*-mutant monkeys revealed a decrease in grey matter but no difference in white matter and cerebrospinal fluid volumes, compared to those of controls (Extended Data Fig. 10a–c).

Altered functional connectivity has recently emerged as a possible biomarker for autism spectrum disorder^{40–42}. We used an unbiased, data-driven, voxel-by-voxel, global and local functional correlation⁴³ (see Methods), and found that long-range connectivity between several brain regions in mutants was reduced relative to that in controls. Notably, we found that in mutants there was hypo-connectivity in putative default-mode networks, including the posterior cingulate cortex, medial prefrontal region and motor regions (Fig. 5a, d–f). We also identified local hypo-connectivity in thalamic and striatal regions (Fig. 5b, c, g, h) in the mutants. By contrast, we detected local hyper-connectivity in the somatosensory cortex, extrastriate cortical areas, and posterior cingulate cortex of the mutant monkeys (Fig. 5b, c, i, j). Seed-based analysis confirmed these results of reduced global connectivity and greater local connectivity in *SHANK3* mutants (Extended Data Fig. 10d–f). Taken together, our MRI data indicate that the mutant macaques have a dysregulated resting-state connectivity, both globally and locally.

Discussion

In this model of the *SHANK3* mutation in macaques, we observed a combination of reduced mobility, increased repetitive behaviours and impaired sociability that reconciles previous studies in rodents^{2,11,26,27}. *SHANK3* mutants exhibit notable sleep disturbances and activity differences, which may assist in the discovery of characteristic biomarkers for Phelan–McDermid syndrome, autism spectrum disorder and other neurodevelopmental disorders in humans^{38,44}. Altered social behaviours and stereotypy (such as licking fingers and cage bars), as well as reduced muscle strength and a lack of vocalization, in the *SHANK3*-mutant monkeys parallels the hypotonia and speech or language impairments of children with autism spectrum disorder or Phelan–McDermid syndrome^{4,8}. An altered pupillary light reflex, such as we found in the mutant monkeys, has been reported in autism spectrum disorder³²; this has not specifically been examined in Phelan–McDermid syndrome, and warrants further exploration. *SHANK3* mutants displayed comparable learning ability to controls in a simple visual discrimination task. However, a more-complex Hamilton search task revealed learning impairments that might arise from reduced flexibility or switching of strategies. Given that we could behaviourally test only a small number of mutants, further cognitive testing of larger groups will be needed to characterize any intellectual impairment.

Autism spectrum disorder is a heterogeneous disorder both in terms of the clinical manifestation of symptoms and its underlying aetiology. In fact, the genetic predisposition for autism spectrum disorder is likely to be different among individuals, even when the same gene is affected^{3–9}. It is therefore difficult to draw general conclusions from a single mutant monkey (as recently reported²⁴), owing to the

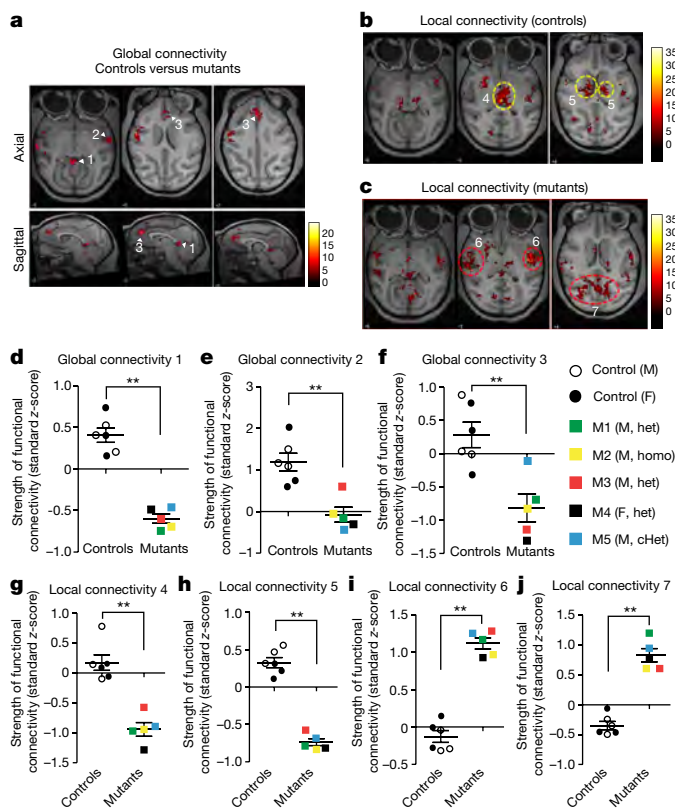


Fig. 5 | Dysregulated global and local connectivity in *SHANK3*-mutant macaques. **a**, Axial and sagittal views of differential global connectivity between control and mutant groups. Clusters with robustly higher global connectivity are highlighted by an arrowhead, and numbered. Putative brain regions are posterior cingulate cortex (1), motor cortex (2) and anterior cingulate cortex (3). **b**, **c**, Axial views of local connectivity in controls and mutants. Clusters with robust alteration of local connectivity are highlighted and numbered with dashed circles. Putative brain regions are thalamic regions (4), striatum (5), somatosensory cortex (6) and near the posterior cingulate cortex and extrastriate cortical areas (7). **d–f**, Normalized global connectivity shows a reduced strength in *SHANK3* mutants for regions 1, 2 and 3 (corresponding to **a**). **g**, **h**, Normalized local connectivity shows a reduced strength in *SHANK3* mutants for regions 4 and 5 (corresponding to **b**). **i**, **j**, Normalized local connectivity shows an increased strength in *SHANK3* mutants for regions 6 and 7 (corresponding to **c**). In **d–j**, $n = 6$ macaques for control group; $n = 5$ macaques for *SHANK3*-mutant group. $P = 0.0043$ (**d**), 0.0087 (**e**, **f**) and 0.0043 (**g–j**). Data are presented as mean \pm s.e.m. Two-tailed Mann–Whitney U -test. Within-group threshold of $P < 0.001$ and $P < 0.05$ family-wise error correction of $P < 0.05$ (with a cluster-forming threshold of $k \geq 25$) were used. Coloured squares indicate individual monkeys with *SHANK3* mutation.

heterogeneity of symptoms that are related to this disorder, as well as the variability of personalities and abilities even in wild-type monkeys. Among the five founder *SHANK3*-mutant macaques, we observed considerable heterogeneity in the severity of behavioural manifestations, including motor impairments, stereotypies and learning problems in a complex task. Differential behavioural outcomes could result from normal inter-monkey variability, different mutation patterns in *SHANK3* and/or different genetic backgrounds owing to the monkeys being outbred. Among the five mutant monkeys, M4 is closer to the wild-type controls than are the other mutants for many behavioural variables (for example, total activity measure; Fig. 2g), which is consistent with its genotype and protein expression. M4 carries an in-frame deletion that causes a reduction in the levels of SHANK3 protein of about 20%; the level of reduction in SHANK3 protein isoforms is greater in the homozygous mutant (M2), the compound-heterozygous mutants (M1 and M5) and the heterozygous mutant (M3). Given the small number

of monkeys in this study, as well as their genetic heterogeneity, all of our findings will need to be confirmed in larger numbers of monkeys in the future. Some of the genetic heterogeneity will be better controlled in F1-generation macaques.

In this initial characterization of *SHANK3* mutants, we are unable to pinpoint any of the causes that might underlie the behavioural differences. Anxiety disorder, exacerbated in a social context, could be a common contributor^{45,46}. The hypoactivity and compromised general exploratory activity that we observed in *SHANK3*-mutant monkeys could contribute to the low reciprocity of social interactions and lack of vocalization, although the movement velocity of the mutants during the social tests did not differ from that of the controls. The altered structural MRI and neural connectivity measures may be immune to possible confounding factors. Clinical studies have shown that children and adults with sleep disorders are prone to develop language problems, and have compromised attention and executive function compared to healthy sleepers⁴⁴. Autistic children with serious sleep disorders may have difficulty controlling repetitive behaviours, and may show a lower performance on tests of attention and memory⁴⁷. The *SHANK3*-mutant monkeys provide an opportunity to test complex interactions such as these in the future.

Autism spectrum disorder is thought to affect multiple interconnected regions of the brain, and there is evidence for alterations in brain connectivity that could contribute to the behavioural phenotypes that are associated with autism spectrum disorder^{39,48}. Recent studies have suggested that non-human primates have a resting-state default-mode network that is similar to that of humans^{49,50}. Our discovery in a non-human-primate model of atypical connectivity in local and long-range circuits—especially in the cingulate, frontal, thalamic and striatal regions—suggests a path for further studies to identify circuit abnormalities and potential biomarkers for treatment studies. Monogenic forms of autism spectrum disorder may also offer insights into altered functional brain connectivity in polygenic or idiopathic autism spectrum disorder. Future longitudinal studies of resting-state functional connectivity, combined with in vivo recordings and circuit manipulations, in the second generation of *SHANK3*-mutant monkeys may allow for an in-depth understanding of the development of aberrant connectivity in neural circuits and their relevance for the behavioural phenotypes that characterize autism spectrum disorder.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-019-1278-0>.

Received: 4 November 2018; Accepted: 13 May 2019;

Published online 12 June 2019.

- Naisbitt, S. et al. Shank, a novel family of postsynaptic density proteins that binds to the NMDA receptor/PSD-95/GKAP complex and cortactin. *Neuron* **23**, 569–582 (1999).
- Jiang, Y. H. & Ehlers, M. D. Modeling autism by *SHANK* gene mutations in mice. *Neuron* **78**, 8–27 (2013).
- Moessner, R. et al. Contribution of *SHANK3* mutations to autism spectrum disorder. *Am. J. Hum. Genet.* **81**, 1289–1297 (2007).
- Phelan, K. & McDermid, H. E. The 22q13.3 deletion syndrome (Phelan–McDermid Syndrome). *Mol. Syndromol.* **2**, 186–201 (2012).
- Betancur, C. & Buxbaum, J. D. *SHANK3* haploinsufficiency: a “common” but underdiagnosed highly penetrant monogenic cause of autism spectrum disorders. *Mol. Autism* **4**, 17 (2013).
- Sanders, S. J. et al. Insights into autism spectrum disorder genomic architecture and biology from 71 risk loci. *Neuron* **87**, 1215–1233 (2015).
- Leblond, C. S. et al. Meta-analysis of *SHANK* mutations in autism spectrum disorders: a gradient of severity in cognitive impairments. *PLoS Genet.* **10**, e1004580 (2014).
- Frank, Y. et al. A prospective study of neurological abnormalities in Phelan–McDermid syndrome. *J. Rare Disord.* **5**, 1–13 (2017).
- Chen, J. A., Peñaarikano, O., Belgard, T. G., Swarup, V. & Geschwind, D. H. The emerging picture of autism spectrum disorder: genetics and pathology. *Annu. Rev. Pathol.* **10**, 111–144 (2015).
- Gauthier, J. et al. De novo mutations in the gene encoding the synaptic scaffolding protein *SHANK3* in patients ascertained for schizophrenia. *Proc. Natl Acad. Sci. USA* **107**, 7863–7868 (2010).

11. Peça, J. et al. *Shank3* mutant mice display autistic-like behaviours and striatal dysfunction. *Nature* **472**, 437–442 (2011).
12. Jennings, C. G. et al. Opportunities and challenges in modeling human brain disorders in transgenic primates. *Nat. Neurosci.* **19**, 1123–1130 (2016).
13. Bauman, M. D. & Schumann, C. M. Advances in nonhuman primate models of autism: integrating neuroscience and behavior. *Exp. Neurol.* **299**, 252–265 (2018).
14. Chang, S. W. et al. Neuroethology of primate social behavior. *Proc. Natl Acad. Sci. USA* **110**, 10387–10394 (2013).
15. Platt, M. L., Seyfarth, R. M. & Cheney, D. L. Adaptations for social cognition in the primate brain. *Phil. Trans. R. Soc. Lond. B* **371**, 20150096 (2016).
16. Izpisua Belmonte, J. C. et al. Brains, genes, and primates. *Neuron* **86**, 617–631 (2015).
17. Sclafani, V. et al. Early predictors of impaired social functioning in male rhesus macaques (*Macaca mulatta*). *PLoS ONE* **11**, e0165401 (2016).
18. Liu, Z. et al. Autism-like behaviours and germline transmission in transgenic monkeys overexpressing MeCP2. *Nature* **530**, 98–102 (2016).
19. Chen, Y. et al. Modeling Rett syndrome using TALEN-edited MECP2 mutant cynomolgus monkeys. *Cell* **169**, 945–955 (2017).
20. Sasaki, E. et al. Generation of transgenic non-human primates with germline transmission. *Nature* **459**, 523–527 (2009).
21. Cong, L. et al. Multiplex genome engineering using CRISPR/Cas systems. *Science* **339**, 819–823 (2013).
22. Niu, Y. et al. Generation of gene-modified cynomolgus monkey via Cas9/RNA-mediated gene targeting in one-cell embryos. *Cell* **156**, 836–843 (2014).
23. Zhao, H. et al. Altered neurogenesis and disrupted expression of synaptic proteins in prefrontal cortex of *SHANK3*-deficient non-human primate. *Cell Res.* **27**, 1293–1297 (2017).
24. Tu, Z. et al. CRISPR/Cas9-mediated disruption of *SHANK3* in monkey leads to drug-treatable autism-like symptoms. *Hum. Mol. Genet.* **28**, 561–571 (2019).
25. Durand, C. M. et al. Mutations in the gene encoding the synaptic scaffolding protein *SHANK3* are associated with autism spectrum disorders. *Nat. Genet.* **39**, 25–27 (2007).
26. Zhou, Y. et al. Mice with *Shank3* mutations associated with ASD and schizophrenia display both shared and distinct defects. *Neuron* **89**, 147–162 (2016).
27. Speed, H. E. et al. Autism-associated insertion mutation (InsG) of *Shank3* exon 21 causes impaired synaptic transmission and behavioral deficits. *J. Neurosci.* **35**, 9648–9665 (2015).
28. Bae, S., Park, J. & Kim, J. S. Cas-OFFinder: a fast and versatile algorithm that searches for potential off-target sites of Cas9 RNA-guided endonucleases. *Bioinformatics* **30**, 1473–1475 (2015).
29. Jiang, Y. & Platt, M. L. Oxytocin and vasopressin flatten dominance hierarchy and enhance behavioral synchrony in part via anterior cingulate cortex. *Sci. Rep.* **8**, 8201 (2018).
30. Falck-Ytter, T., Bölte, S. & Gredebäck, G. Eye tracking in early autism research. *J. Neurodev. Disord.* **5**, 28 (2013).
31. Mosher, C. P., Zimmerman, P. E. & Gothard, K. M. Videos of conspecifics elicit interactive looking patterns and facial expressions in monkeys. *Behav. Neurosci.* **125**, 639–652 (2011).
32. Daluwatte, C. et al. Atypical pupillary light reflex and heart rate variability in children with autism spectrum disorder. *J. Autism Dev. Disord.* **43**, 1910–1925 (2013).
33. Maestripietri, D. & Wallen, K. T. Affiliative and submissive communication in rhesus macaques. *Primates* **38**, 127–138 (1997).
34. Hinde, R. A. & Rowell, T. E. Communication by postures and facial expressions in the rhesus monkey (*Macaca mulatta*). *J. Zool.* **138**, 1–21 (1962).
35. Gothard, K. M., Battaglia, F. P., Erickson, C. A., Spitzer, K. M. & Amaral, D. G. Neural responses to facial expression and face identity in the monkey amygdala. *J. Neurophysiol.* **97**, 1671–1683 (2007).
36. Parr, L. A. & Heintz, M. Facial expression recognition in rhesus monkeys, *Macaca mulatta*. *Anim. Behav.* **77**, 1507–1513 (2009).
37. Wass, S. V. et al. Shorter spontaneous fixation durations in infants with later emerging autism. *Sci. Rep.* **5**, 8284 (2015).
38. Tabet, A. C. et al. A framework to identify contributing genes in patients with Phelan-McDermid syndrome. *NPJ Genom. Med.* **2**, 32 (2017).
39. Rudie, J. D. et al. Altered functional and structural brain network organization in autism. *Neuroimage Clin.* **2**, 79–94 (2013).
40. Emerson, R. W. et al. Functional neuroimaging of high-risk 6-month-old infants predicts a diagnosis of autism at 24 months of age. *Sci. Transl. Med.* **9**, eaag2882 (2017).
41. Lewis, J. D., Theilmann, R. J., Townsend, J. & Evans, A. C. Network efficiency in autism spectrum disorder and its relation to brain overgrowth. *Front. Hum. Neurosci.* **7**, 845 (2013).
42. Buckner, R. L., Andrews-Hanna, J. R. & Schacter, D. L. The brain's default network: anatomy, function, and relevance to disease. *Ann. NY Acad. Sci.* **1124**, 1–38 (2008).
43. Whitfield-Gabrieli, S. & Nieto-Castanon, A. Conn: a functional connectivity toolbox for correlated and anticorrelated brain networks. *Brain Connect.* **2**, 125–141 (2012).
44. Goldman, S. E. et al. Defining the sleep phenotype in children with autism. *Dev. Neuropsychol.* **34**, 560–573 (2009).
45. Adolphs, R. The social brain: neural basis of social knowledge. *Annu. Rev. Psychol.* **60**, 693–716 (2009).
46. Arnsten, A. F. Stress signalling pathways that impair prefrontal cortex structure and function. *Nat. Rev. Neurosci.* **10**, 410–422 (2009).
47. Guérolle, F. et al. Melatonin for disordered sleep in individuals with autism spectrum disorders: systematic review and discussion. *Sleep Med. Rev.* **15**, 379–387 (2011).
48. Just, M. A., Keller, T. A., Malave, V. L., Kana, R. K. & Varma, S. Autism as a neural systems disorder: a theory of frontal-posterior underconnectivity. *Neurosci. Biobehav. Rev.* **36**, 1292–1313 (2012).
49. Moeller, S., Nallasamy, N., Tsao, D. Y. & Freiwald, W. A. Functional connectivity of the macaque brain across stimulus and arousal states. *J. Neurosci.* **29**, 5897–5909 (2009).
50. Vincent, J. L. et al. Intrinsic functional architecture in the anaesthetized monkey brain. *Nature* **447**, 83–86 (2007).

Acknowledgements We thank L. Harp McGovern and the late P. J. McGovern for their vision and support; F. Zhang for advice and reagents for CRISPR; D. G. Amaral for sharing image resources for creating eye-tracking stimuli; J. Bachevalier for advice on behavior testing; E. A. Murray for guidance on the Wisconsin General Test Apparatus assay; G. Genovese and R. Rosario for support with statistical and bioinformatics analysis; S. Sharma, S. Lall and S. Krol for critical reading of the manuscript; L. Dennis, N. Nien-Chu Espinoza, S. Yang, A. Chakrabarti, N. Joshi and Y. Fukumura for behavioral scoring; X. Wu, X. Ding, L. Cheng and X. Liu for technical support; the veterinary team of Blooming-Spring for excellent colony management and technical support; and S. E. Hyman (Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard), N. Sanjana (NYU) and L. Cong (Stanford University) and members of the Feng laboratory at MIT for critical discussion on this project. This work was supported by National Key R&D Program of China (2017YFC1307500); Shenzhen Overseas Innovation Team Project (KQTD20140630180249366); Guangdong Innovative and Entrepreneurial Research Team Program (2014ZT05S020). S.Y. and Q.K. was supported by Frontier and Innovation of Key Technology Project in Science and Technology Department of Guangdong Province (2014B020225007 and 2019B020235002); and Program for New Century Excellent Talents in University of Ministry of Education of the People's Republic of China (NCET-12-1078). This work was also supported by the National Key R&D Program of China (2018YFA0107203 and 2017YFA0103802 to A.P.X., 2017YFA0103802 to W.L.); the External Cooperation Program of Chinese Academy of Sciences (172644KYSB20160026); International Partnership Program of Chinese Academy of Sciences (172644KYS820170004 to L.W., 172644KYSB20160175 to H.Z.); the Patrick J. McGovern Foundation; Hundred Talent Program of Chinese Academy of Sciences to H.Z.; the National Natural Science Foundation of China (81425016 to A.P.X., 31671119 to Z.L.); Shenzhen Science and Technology Innovation Commission grants (JCYJ20151030140325151 to H.Z.; GJHZ20160229200136090, JCYJ20170413165053031 to T.Y.; JCYJ20170413162938668 to Z.L.). Y. Zhou was supported by postdoctoral fellowships from the Simons Center for the Social Brain at MIT and Nancy Lurie Marks Family Foundation. G.F. is supported by the McGovern Institute for Brain Research at MIT, James and Patricia Poitras Center for Psychiatric Disorders Research at MIT, the Stanley Center for Psychiatric Research at the Broad Institute of MIT and Harvard, the Hock E. Tan and K. Lisa Yang Center for Autism Research at MIT, and Edward and Kay Poitras. L.W. is also supported by Guangdong Provincial Key Laboratory of Brain Connectome and Behavior 2017B030301017, Shenzhen Discipline Construction Project for Neurobiology DRCSM [2016]1379, and Shenzhen-Hong Kong Institute of Brain Science.

Reviewer information *Nature* thanks Thomas Bourgeron, Michael Platt and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions G.F., S.Y. and Y. Zhou conceived the study, and R.D., G.F. and H.Z. provided ongoing guidance on the design. S.Y. and A.P.X. oversaw the generation of mutant monkeys. H.Z. oversaw the characterization of mutant monkeys. Y. Zhou carried out CRISPR design and validation. S.Y., Q.K., H.C., Y. Zhou, J.Y., D.X., Y.H. and A.P.X. generated mutant monkeys. Y. Zhou and D.W. designed and performed molecular, protein, sequencing and off-target analyses. R.L., J.S., Y. Zhou, G.F. and R.D. designed and analysed behavioural experiments and MRI assays. H.Z., L.W., Z.L., T.Y., Y. Zou, M.J., W.J., Y.B., W.M., T.A., Y.L., X.L., W.L., L.H., S.A.A. and M.S. participated in the design or execution of some of the behavioural experiments. R.L., D.S.H., J.W.F. III, J.B.H., A.F.-K., O.M. and S.P. managed and performed behavioural scoring. Y. Zhou, R.L., R.D., J.S. and G.F. wrote the manuscript with input from all authors.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-019-1278-0>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-019-1278-0>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to H.Z., A.P.X., G.F. or S.Y.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

METHODS

CRISPR construct, cell culture and mRNA preparation. Selected gRNAs that target exon 21 of macaque *SHANK3* were cloned into the pU6-gRNA expression plasmid separately, using the two BbsI restriction sites. Sequence-verified gRNA-expressing plasmids were individually co-electroporated into primary cultured fibroblast cells together with wild-type *SpCas9*-expressing plasmid (Primate Biologicals) and cultured for 48 h using dermal fibroblast culture medium (Zenbio). Cells were collected, and total genomic DNA was extracted using a NucleoSpin tissue kit (Macherey-Nagel). An amplicon of 700 bp flanking the targeted mutation region of *SHANK3* was amplified from each DNA sample, followed by analysis using a standard surveyor nuclease-based mutation detection assay as described in the user manual (IDT DNA). To prepare CRISPR mRNAs for embryo injection, guide RNA no. 1 and no.2, and wild-type *SpCas9*-expressing plasmid, were linearized and transcribed with commercial kits (MEGAscript T7 (Ambion) and HiScribe T7 (NEB)). Synthesized mRNA was further purified with MEGAclear Transcription clean-up kit (Ambion) and quantified with Nanodrop 2000 (Thermo Fisher). Monkey embryos injected with *SHANK3* CRISPR mRNA were cultured for 5 to 7 days and collected individually, lysed with protease K; this was followed by nested PCR-based genotyping of 300-bp amplicons using two sets of primers. Primer set I forward: 5'-gaccctgtttgtggtatgacagg-3'; reverse: 5'-cctggccggggtctgtgggtgccc-3'; primer set II forward: 5'-ggctcctggctccggccttc-3'; reverse: 5'-gtgggcagggtcctgcccacagtcg-3'. The presence of indels in the injected embryos was revealed by electrophoresis-based separation on 4% agarose gel and subsequent Sanger sequencing of the PCR product.

Genotyping of mutant monkeys and off-target analysis. To genotype *SHANK3* mutation in newborn macaques, a 1-kb PCR product flanking the desired mutation site was amplified using high-fidelity polymerase (PrimeSTAR HS DNA Polymerase with GC buffer) from genomic DNA of each monkey, using primer forward: 5'-atcgattcgctccactgtgtctgtcgt-3' and reverse: 5'-tagggatcatgtgtctgtggtctccagggtggggc-3'. PCR products from individual monkeys were cloned separately into pBSKII vector using the EcoRI and BamHI restriction sites. Ligation mixtures were transformed into competent *Escherichia coli* and plated onto plates with ampicillin selectivity. Individual bacterial colonies were picked up, and plasmid DNA was subjected to Sanger sequencing to examine various types of indels and their percentages. Possible off-target sites caused by administration of single-guide (sg)RNA1- and sgRNA2-guided Cas9 endonucleases for exon 21 of *SHANK3* gene were predicted by scanning *Macaca fascicularis* 5.0 of the UCSC genome database through Cas-OFFinder, a bioinformatics-based algorithm (<http://www.rgenome.net/cas-offinder/>). Genomic DNA from peripheral whole blood from all five *SHANK3* mutants and two wild-type monkeys was extracted and used for subsequent analysis. The top ten predicted off-target sites ranging from one to three mismatches within targeted sequences of each sgRNA were amplified by a high-fidelity polymerase. Purified PCR products of each predicted site from individual monkeys were aligned and examined using Sanger sequencing. Detailed annotation of all predicted off-target sites, primers used for PCR amplification and summary of sequencing results are available in Supplementary Table 1.

Craniotomy and brain biopsy. Surgical procedures were performed after completing all behavioural analysis and MRI scanning presented in this study. While the monkeys were under general anaesthesia, a craniotomy was performed and a 2-mg biopsy was taken from the surface of V1.

Western blotting and semi-quantification of protein expression. Biopsy samples from the superficial V1 layer of each monkey were briefly sonicated inside 100 µl of PBS solution containing protease inhibitor (Roche). A sonication protocol consisting of 10% power and three pulses with 30% on and 70% off (Omni-Ruptor 250) was carried out to ensure complete lysis of tissue sample. After quantification with BCA kit (Pierce), the calculated amount of each protein was mixed with an equivalent amount of 2× Laemmli sample buffer (Bio-Rad) and boiled for 5 min at 95°C. Sample volumes corresponding to 20 µg of total protein amount per lane were loaded onto 4–15%-gradient Mini-PROTEAN TGX gels (Bio-Rad) and ran for 3 h at 60 V. The proteins were then transferred onto Whatman Protran nitrocellulose membranes (0.2-µm pore size, BA83, Sigma Aldrich) using a tank blot system (Mini Trans-Blot Cell, Bio-Rad) for 120 min at 100 V at 4°C. The membranes were blocked for 1 h with 5% non-fat milk dissolved in TBS buffer that did not contain any Tween-20. Subsequently, the membranes were incubated with primary antibodies diluted in Odyssey blocking buffer (Li-COR Biosciences) with a dilution factor of 1:100 for N-terminal *SHANK3* antibody (NIH NeuroMab N367/62) and 1:500 for C-terminal *SHANK3* antibody (Santa Cruz Biotech SC-30193) for 12 h at 4°C. Following primary antibody incubation, the membranes were washed 3 times for 5 min per wash using TBST buffer (0.05% Tween-20). Then, the secondary antibodies, goat-anti-mouse IRDye 680 (Li-COR Biosciences), donkey-anti-rabbit IRDye 800CW (Li-COR Biosciences) diluted in 1:1 TBST (0.05% Tween-20):Odyssey Blocking Buffer (Li-COR Biosciences), were incubated with the membrane for 1 h at room temperature. Following three rounds of washing with TBST, the membranes were scanned using an Odyssey CLx infrared

imaging system (Li-COR Biosciences). Specific bands were then quantified with the contrast-independent, automatic background-subtraction rectangular region-of-interest (ROI) tool of the built-in Software Image Studio 3.1 (Li-COR Biosciences), and normalized to an α -tubulin loading control for each lane and each blot. The values obtained for each sample from mutant monkeys were then normalized to the wild-type expression. To estimate the percentile of remaining isoforms of the *SHANK3* protein, five technical repeats of western blot using the same protein lysis were analysed and averaged values for each mutant monkey were presented.

Statement on animal work. All animal-related work was done in accordance with NIH guide for the care and use of laboratory animals (<https://www.ncbi.nlm.nih.gov/books/NBK54050/>) and institutional animal care and use guidelines approved by the IACUC of Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences. We also followed nc3r recommendations (<https://www.nc3rs.org.uk/>) by using the minimum number of mutants and age-matched controls, while maintaining statistical reliability. The cynomolgus monkey (*M. fascicularis*) breeding facilities, housing and primate laboratories used in this study are accredited by The Association for Assessment and Accreditation of Laboratory Animal Care. Wild-type and mutant monkeys in the experimental groups were housed in an environmentally controlled facility (temperature: $22 \pm 1^\circ\text{C}$, humidity: $50 \pm 5\%$ relative humidity) with 12-h light/12-h dark cycle (lights on at 07:00). All macaques were fed with commercial monkey diet twice per day, plus one meal of seasonal fruits daily, and with free access to a water bottle. Monkeys were under careful veterinary monitoring twice per day to evaluate and ensure their health status. Infant monkeys were housed with their mothers and weaned at 12 months of age. After weaning, pairs of young monkeys were selected for compatibility and pair-housed together in both the wild-type control group and the *SHANK3*-mutant group. The pairing strategy for the mutant group was: M1 with M3, and M2 with M5; the female mutant M4 was paired with a female from the control group. A sliding divider was used to temporarily separate two monkeys during certain procedures, including video recording of home-cage activity, the wearing of motion watches and recovery from anaesthesia or surgical procedures.

Surgical procedures and generation of *SHANK3*-mutant monkeys. Detailed procedures related to superovulation, oocytes collection, intracytoplasmic sperm injection (ICSI), CRISPR injection, embryo transfer, and pregnancy diagnosis are documented as previously reported⁵¹. In brief, female cynomolgus monkeys aged 7–12 years with regular menstrual cycles were chosen as oocyte donors for superovulation and recipients for embryo transfer after genetic engineering. Healthy male monkeys aged 5–10 years of proven fertility were chosen as sperm donors. Under general anaesthesia, laparoscopic procedures were performed aseptically for both oocyte collection and embryo transfer in females. Female monkeys were intramuscularly injected with recombinant human FSH (rhFSH) (EMD Serono) twice per day consecutively for eight days, followed by one injection of recombinant human chorionic gonadotropin (rhCG) (EMD Serono) on day 9. Oocytes were collected through laparoscopic-guided follicular aspiration 33–36 h after administration of rhCG. Mature oocytes at stage MII were selected to perform fertilization using ICSI. Seven-to-eight hours after sperm injection, embryos with the clear appearance of two pronuclei were injected into cytoplasm with a mixture of mRNAs containing 50 ng/µl of *SpCas9*, 25 ng/µl of gRNA 1 and 25 ng/µl of gRNA 2. Injected embryos were immediately transferred into the oviduct of the stage-matched recipient female monkeys. Successful implantation of embryos and pregnancy of recipient were examined by ultrasonography four weeks after embryo transfer.

Data analysis and statistical comparisons. All behavioural observations, scoring, data analysis in this study were carried out by trained observers without previous knowledge of the experimental design and goal of this study. All data are presented as mean \pm s.e.m. Comparisons between *SHANK3*-mutant monkeys and the wild-type control group were analysed using a nonparametric, two-tailed Mann–Whitney *U*-test, unless otherwise specified in the legend of each figure (GraphPad Prism 5). * $P < 0.05$; ** $P < 0.01$; *** $P < 0.001$. Statistical power was calculated using the following steps. First, statistical tests were run on the control and mutant measurements; sample means and standard deviations were then calculated; Cohen's *d* effect size from two distributions with the empirical sample means and standard deviations were then computed. Finally, simulations were run to generate a nominally significant *P* value from two distributions with the empirical sample means and standard deviations (see Supplementary Table 3 for summary of power calculations).

Activity and sleep monitoring. Activity and sleep patterns of the monkeys were recorded after suitably modifying a commercially available actigraphy device with a built-in ambient light sensor and event marker as previously described⁵². In brief, monkeys were lightly anaesthetized with ketamine, and a motion watch was placed individually. Monkeys were habituated to wear the watch for two weeks before data collection. Following habituation, daily physical activity and a full range of sleep

and circadian parameters for each monkey were recorded and collected for up to seven days. Recorded data were transferred to a host computer and analysed.

Behavioural scoring. The behaviour of the monkeys during the first year of life (starting at one month of age) was scored from video recordings (Canon, 1,920 × 1,080 at 30 frames per second) using Observer software (Noldus Observer XT 11). During this time, the monkeys were housed together with their mother. Each observer was trained by an expert observer on a subset of videos until inter-observer reliability with a highly experienced observer exceeded 0.85. After that, each observer scored a subset of the videos for data collection. Per monkey, between 12 and 49 sessions of 10 min in length each, spread throughout the year, were scored, depending on the availability of video material. The sessions were binned into periods of 16 weeks to ensure the equal contribution of periods with different sampling density to the mean for the year. We used roaming entropy to quantify the locations that animals visited in the home cage. The location of the animals in the cage was annotated as being on the left wall, right wall, floor, front wall, rear wall or ceiling, at each point in time. To quantify the evenness of the distribution of locations for each animal, entropy was calculated using the proportions of time spent in each location as previously described⁵³. As an example, a monkey that spends all of its time in one location has an entropy of 0, and a monkey that spends an equal amount of time in all locations has maximum entropy, 2.585 in the present case of 6 possible locations. The coding of stereotypy of mutants included flipping, licking cage bars and licking fingers. The unit of measurement is time performing the behaviour in seconds, per minute of observation.

Habituation, social interaction and exploration. Habituation of monkeys and social interaction with other monkeys was tested in a room and cage separate from the home room and housing cage. The cage dimensions were 3 m × 1.5 m × 1.5 m (width × height × depth). Each monkey was habituated to the novel environment separately, by placing them inside the cage for 30 min on 2 separate days. A video camera was used to record the habituation and social interaction sessions. Each social interaction session consisted of the following procedure: two monkeys were taken out of their home cage, transported to the social interaction room and placed inside the cage. The monkeys were separated by a vertical divider. The divider was made of non-transparent plastic material, which blocked visual and tactile communication but not auditory and olfactory contacts between two monkeys. After approximately 10 min, the divider was removed and the monkeys could interact for 30 min. After this, the divider was re-inserted. Finally, after another 10 min, the monkeys were taken back to the home cage. Each animal in our experimental group (six controls and five mutants) went through ten social interaction sessions, during which the monkey met ten age-matched wild-type monkeys (the probe group) in random order.

Behaviour scoring during social interaction. The video material was scored for categories of behaviours, including chase, flee, follow, circle, play, attack, presentation of rump, anogenital inspection, groom and mount. The behaviours chase, follow, flee, circle and play were combined into an aggregate category labelled ‘social behaviours.’ We report the average total duration per 5 min for each category in the main text, and component categories in Extended Data Figs. 7, 9.

Movement tracking during habituation and social interaction. Computer vision was used to track the animals in two dimensions (x and y) during the habituation and social interaction sessions. Although actual movement occurs in three dimensions, we assume that there was enough information in the two-dimensional plane to detect potential differences between mutant and controls (see diagram in Extended Data Fig. 8f).

Tracking is performed according to the Bayesian nonparametric, linear dynamic system. $y_n^t \in \mathbb{R}^5$ is the n th observation at time t , and denotes a (u, v, L, a, b) tuple of pixel coordinates (u, v) and colour (L, a, b) . The observation model is a per-observation Gaussian mixture

$$p(y_n^t | z_n^t, b_n, \theta_p^t) = N(y_n^t | b_n, \Sigma_B) \prod_p N(y_n^t | \mu_p^t, \Sigma_p^t) \mathbb{I}(z_n^t=0) \mathbb{I}(z_n^t=p)$$

in which $\theta_p^t = (\mu_p^t, \Sigma_p^t)$ are Gaussian parameters that can be viewed conceptually as corresponding to object-part (for example, body, tail or head) locations and appearance (μ_p^t) , and extents (Σ_p^t) of the objects to be tracked, and $b_n \in \mathbb{R}^5$ is a background model for the n th observation—also a (u, v, L, a, b) tuple. Part locations and appearances are generated as

$$p(\mu_p^t | \delta_p^t, x_k^t) = \text{Unif}(\mu_p^t | U_X) \prod_k N(\mu_p^t | Hx_k^t, \Sigma_X) \mathbb{I}(\delta_p^t=k)$$

The z_n^t variables are categorical random variables with a Griffiths–Engen–McCloskey prior; they assign observations to either background or object parts. The δ_p^t variables are also categorical random variables, but with a Dirichlet prior, as we assume a known number of targets. They assign part θ_p^t to one of K object, or to a (uniform) clutter distribution. Matrix $H = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$ projects the

latent object location into the observation space. The μ_p^t and Σ_p^t components share a data-dependent normal-inverse-Wishart (NIW) prior at each time t with concentration $\frac{1}{N_p} \sum_n y_n^t$ and scale $\sum_n y_n^t (y_n^t)^T \mathbb{I}(z_n^t = 1)$ in which $N_p = \sum_n z_n^t$ and z_n^t is a categorical random variable sampled such that $p(z_n^t = 0) \propto N(y_n^t | b_n, \Sigma_B)$ and $p(z_n^t = 1) \propto \text{Unif}(y_n^t | U_X)$. Latent x_k^t variables are the independent, time-evolving image location and velocity of each tracked object with dynamics

$$p(x_k^{t+1} | x_k^t) = N(x_k^{t+1} | Ax_k^t, \Sigma_X)$$

in which $A = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$ enforces a random acceleration model. The inference proceeds as: for each n , estimate background model b_n as median $(y_n^{1:T})$, which can be performed in $O(\log(T) \log(\log(T)))$ time with a randomized algorithm.

Then, for time $t = 1, \dots, T$ observation $n = 1, \dots, N$. (1) Sample z_n^t , then estimate NIW priors for θ_p^t ; (2) estimate $z_n^t, \mu_p^t, \Sigma_p^t$ using variational approximation⁵⁴; and (3) propagate x_k^{t-1}, P_k^{t-1} to \hat{x}_k^t, \hat{P}_k^t according to dynamics, then sample $\delta_p^t | \mu_p^t, x_k^t$ and compute Kalman filtered estimate $\bar{x}_k^t, \bar{P}_k^t | \{\mu_p^t : \delta_p^t = k\}$ in which \bar{P}_k^t is the covariance estimate of random variable x_k^t with mean \bar{x}_k^t (ref. 55). For time $t = T, \dots, 1$, compute smoothed estimates x_k^t via Rauch–Tung–Striebel smoothing.

Sample object trajectories can be computed as Gibbs iterations of the above, although even a single pass over the data yields good performance under reasonable conditions. Being generative, this model provides several benefits over more-standard, detection-based trackers. The output of the computer vision tracker used for the present analysis are—for each video frame—the coordinates of the centre of gravity of the pixels associated with each monkey, as well as velocity. To account for small variations in camera position and zoom level, the tracking data were normalized such that for x , 0 represents the right cage edge and 1 the left, and for y , 0 is the cage floor and 1 the top. **Location preferences.** Movement during the last habituation session was analysed to determine the location preferences in control monkeys and mutants. The proportion of time spent in the left half versus the right half, and upper half versus lower half were determined using x and y location values. Here, a value of 0.5 means that the monkeys were unbiased in their exploration of locations.

Eye tracking. The eye movements of the monkeys were examined while they observed images and video clips, using a Tobii TX300 eye tracker (sampling frequency 120 Hz) under semi-freely moving conditions, without head restraint. For a subset of sessions, the monkeys were tested while in a specially designed ‘eye-tracking box’ with a view-port cut out. For another subgroup of sessions, the monkeys were seated in a typical primate chair. The eye tracking box is a 50-cm × 50-cm × 50-cm box made of ½-inch-thick opaque plexiglass. A 1-inch × 3-inch viewport cut out allowed the monkey to see outside (see diagram in Extended Data Fig. 8g). The monkeys were habituated to being in the box by successive exposures that increased in duration. The monkey was lured into the box with fruit. The animals received their typical diet, and were not water- or food-restricted. The eye tracker was calibrated for each monkey separately using built-in Tobii calibration routines.

Separate sessions were run with video clips as stimuli and with still images as stimuli. Video clips were played without audio. A large proportion of the video and still stimuli were created and shared by the laboratory of D. Amaral (with permission).

Video stimuli. For video sessions, 316 video clips (1,920 × 1,080 pixels) of 5–10 s in duration each were created in the following categories.

Whole body. Individual rhesus monkeys in a cage, subcategorized by emotional expression, behaving in a neutral, aggressive or submissive manner towards the camera.

Close up. Individual rhesus monkeys in a cage, behaving in a neutral, aggressive or submissive manner towards the camera. Close-ups were created by cropping original footage centred on the head and resized to 1,920 × 1,080 pixels resolution.

Group. Clips of groups of rhesus monkeys engaged in foraging, mating, grooming, aggression, play and neutral, non-social behaviour.

Miscellaneous. A mixture of 5–12-s-long clips with either abstract moving images, scenes from wildlife movies with non-primate mammalian species (farm animals, birds and dogs), children’s TV shows, computer games, silent films, popular music videos and motion pictures. The 316 videos were split into three sets of 10–12 min, each containing all categories. There was a 1-s black screen interval between subsequent videos. We showed one set per session and ran one session per day. This test was repeated three times. The first time, monkeys were in the eye-tracking box. Thus, there was a total of 9 video-watching sessions with eye tracking, and each of the 316 video clips was shown 3 times.

Video eye-tracking analysis. Using Tobii Studio software, ROIs were drawn around the body, head, eyes and mouth in the whole body, close up and group videos. Eye fixations were detected using the Tobii I-VT filter at the default setting. Data were exported from Tobii studio to .csv format and analysed with custom-written MATLAB and Python scripts. The fraction of time the monkey spent watching each video clip was calculated: Fraction watched = number of samples in which the monkey was gazing within the area taken up by the video/number

of samples during which the clip was shown. In addition, the latency of pupillary light reflex to the onset of the videos was calculated. The pupil diameter was derived from every sample and averaged between left and right eye. Pupil diameter in the period between the start of each video clip, and 3.3 s into the video, was averaged across all videos for each subject. The light reflex magnitude is the difference between the diameter at $t = 0$ and the minimum diameter in the 3.3-s period. After fitting a spline to the curve, the latency of the pupillary light reflex is the point at which the diameter has reduced 0.15 of the light reflex amplitude. During some sessions, the monkey was recorded using a camcorder during eye tracking to study the animal's overt facial expressions (in particular, lip smacking). A trained observer (S.P.) annotated the video using Noldus Observer software, to determine the amount of time spent lip smacking during each video.

Still images. Still images were used in two sessions. In both, a series of images were shown, in which two photographs were shown side-by-side. The 'face versus object' session contained 44 images, each one pairing a photograph of a neutral monkey face or portrait with a photograph of plant leaves. The images of plant leaves are from the MSRC image database (<http://research.microsoft.com/vision/cambridge/recognition/>). In the 'neutral versus threat' session, 31 images were shown, each one pairing two portraits of the same individual monkey (one with a neutral facial expression and one with an open mouth threat expression). In both sessions, each individual monkey appeared in only one image. In most of the portraits, the monkey in them was directly looking at the camera. ROIs were drawn around the eyes and mouth, and around each picture as a whole.

Wisconsin General Test Apparatus tests. The Wisconsin General Test Apparatus (WGTA) was modelled after the original description of the apparatus⁵⁶. The apparatus is a cage where the monkey has access to a presentation tray with four food wells, a trial door, an access door to separate the subject and presentation tray, and a camera for recording. All tests were carried out in a quiet and standardly lit room. Each trial began with opening the access door to give the monkey the chance to reach and open a food well on the presentation tray (which was counted as a response). Once a response was made, the sliding opaque door was lowered, the response was scored as correct or incorrect, and the next trial was set up. Monkeys M2 and M5 were not able to perform any of the WGTA tasks. Therefore, the data only include three mutants. To motivate the monkeys for these tasks, they were not fed before the testing session that day. Throughout the WGTA and Hamilton tasks, six wild-type monkeys from the control group and three *SHANK3*-mutant monkeys underwent the same brief fasting regime each day before their trials.

Adaptation. Before testing, the monkeys were adapted to the testing apparatus, including training to take food rewards (such as fruit) from open food wells. Once the monkeys were able to take the rewards, they were trained to displace a removable disc covering the food wells to obtain the reward. Once the monkeys were able to reliably obtain food rewards by removing objects from the food wells (23 correct out of 25 trials per day), they were trained to perform other tasks.

Black-white discrimination and reversal. Experiments were designed as previously described⁵⁷. In this test, monkeys were required to associate a colour (a black or white block) with a food reward (discrimination phase). Once monkeys learned this association, the opposite colour was rewarded, requiring them to reverse their strategy (reversal phase). Monkeys received 25 trials per day, 5 days per week, until a set criterion for 90%-correct responses was met.

Hamilton search task. The Hamilton search task was based on previous descriptions^{57,58}. In this test, a monkey sat and faced four identical wells with lids. One of the wells contained a food reward and the subject's task was to locate the food with the least number of box openings. In the first stage of the task (the 'set making' stage), the monkeys developed a search strategy to locate a food reward placed randomly in one of four wells with lids, with the exception being that the same location was never rewarded on two consecutive trials. Monkeys are required to lift the lid on the well to obtain the reward. A trial was a sequence of responses that ended with the monkey finding the food. Here the optimal solution is one in which the last well that contained food is avoided, and all other wells are opened only once. On each trial, the experimenter recorded which wells the monkey tried and in what order. The trial was terminated when the subject opened the correct well or when 60 s was up, whichever came first. The criterion was reached when the test subject successfully completed five trials in a row. Monkeys received 25 trials per day for 5 consecutive days.

In the second stage (set breaking), the well that was least-preferred in stage one was now baited on every trial, and the animal was allowed to open as many wells as necessary to locate the reward. The well that was baited did not change across trials within a test session. A perfect strategy would be to quickly discern that the same well location is always rewarded, and to only open that well on every trial. Monkeys received 25 trials per day for 5 consecutive days. We counted the number of times the monkey opened the correct well in the first try (X). For each monkey, we report the difference of X ('delta') between fifth and first day as a measure of learning.

The third stage (forced set breaking) was identical to stage two, but with a different rewarded location from stage two, and with the exception that only one well

opening was allowed on every trial. The single response on that trial was marked correct or incorrect. Monkeys received 25 trials per day for 5 consecutive days. We report the percentage of correct trials for each of the five days.

Structural and functional MRI. Protocols for anaesthesia and MRI were in accordance with NIH guidelines, and were reviewed and approved by the Institutional Animal Care Committee of SIAT. In preparation for MRI scanning under anaesthesia, the monkey was premedicated with atropine (0.05 mg/kg, intramuscular) to decrease bronchial secretions, followed by ketamine (15 mg/kg, intramuscular). To prevent hypothermia, the monkey was wrapped in a blanket, the extremities were covered with mittens and were placed on a warm-water circulating blanket. The anaesthesia was maintained for the duration of the scan with intravenous propofol by continuous infusion using a syringe pump at the dose of 0.5 mg/kg/min. The anaesthetic level was adjusted to eliminate movement as assessed by toe pinches, while keeping the heart rate in the range of 100–140 beats per minute. Corneal reflexes were consistently absent. Electrocardiography, heart rate and oxygen saturation (SpO_2) (range 94–100%) were continuously monitored with an magnetic-resonance compatible monitoring system (Invivo). Rectal temperature was continuously monitored and was maintained between 37.5–38.5 °C. Magnetic resonance scanning was performed at the SIAT imaging centre on a 3T Tim Trio scanner (Siemens) using a custom-designed 8-channel radio-frequency surface head coil. Functional MRI data were acquired using a $T2^*$ -weighted gradient-echo echo-planar sequence (volume repetition time (TR) = 2.1 s, $T2^*$ echo time (TE) = 25 ms, flip angle = 90°, 1.25 mm × 1.25 mm in plane resolution and slice thickness = 1.3 mm). In all monkeys, the slices were acquired using contiguous, interleaved acquisition; 128 volumes per run (3 consecutive runs, each lasting approximately 5 min) were acquired in each monkey. $T1$ -weighted, magnetization-prepared rapid gradient echo structural images were also acquired (TR = 2.1 s; TE = 3.21 ms, flip angle = 8°, 0.5 mm isotropic voxels) and used to align the functional data to the monkey atlas as previously reported⁵⁹.

Functional MRI data analysis. Resting-state functional MRI (rsfMRI) data were first preprocessed in SPM8 (Wellcome Department of Imaging Neuroscience; www.fil.ion.ucl.ac.uk/spm/), using standard spatial preprocessing steps. Data were motion-corrected, realigned, normalized to structural scans (using the monkey atlas template) and spatially smoothed with a 3-mm full-width at half maximum Gaussian kernel. The structural image of each monkey was segmented into white matter, grey matter and cerebral spinal fluid, using SPM8 as previously described⁶⁰.

Motion artefact detection. We used a standard artefact rejection toolbox (www.nitrc.org/projects/artifact_detect/) for a comprehensive analysis of sources of artefacts in time-series data, including spiking and motion to identify outlier data points (TR values), defined as volumes that exceeded three z -normalized standard deviations away from mean global brain activation across the entire volume, or a composite movement threshold of 0.5 mm scan-to-scan frame-wise displacement. There was no significant difference in the number of outliers between groups. In addition, there was no significant difference in the mean or maximum head motion parameters, or the mean or maximum global signal change between groups. Maximum motion (controls 0.09 ± 0.01 versus mutants 0.099 ± 0.01 ; $P = 0.66$); mean motion (controls 0.036 ± 0.01 ; mutants 0.036 ± 0.01 ; $P = 0.48$).

Connectivity analysis. Functional connectivity analysis of rsfMRI data was carried out in custom software developed in MATLAB toolbox (www.nitrc.org/projects/conn/). To minimize partial volume effects with adjacent grey matter, the white matter and cerebral spinal fluid masks were eroded by one voxel and used as a noise ROI. The first three principal components of the signals from the eroded white matter and cerebral spinal fluid were regressed out through an anatomical component-based noise correction approach (aCompCor) as previously described⁶¹. In the aCompCor approach, segmented white matter and cerebral spinal fluid masks are eroded by one voxel to produce the white matter and cerebral spinal fluid noise ROIs. The erosion removes about 70% of the white matter voxels and 95% of the cerebral spinal fluid voxels from the original segmentations. Therefore, unlike the global signal regression approach, the white matter and cerebral spinal fluid masks used in the analysis were only a very small fraction in size, compared to the whole brain mask, which minimizes spurious correlation values. To minimize head-motion-related confounding factors, realignment parameters and their first-order derivatives, along with the motion outliers, were also regressed during de-noising. A temporal band-pass filter of 0.0025–0.05 Hz was applied to the pre-processed functional dataset. For resting-state local and global functional connectivity analysis, we used an unbiased data-driven approach instead of the standard seed-based analysis.

Integrated local correlation. Integrated local correlation (ILC) was used to assess the coupling of local neuronal processes in any given voxel's neighbourhood. The coupling of local neuronal processes influences coherence in a voxel's neighbourhood. The ILC for each voxel is an integration of its spatial correlation function. In this approach, physiological fluctuations owing to respiratory and cardiac fluctuations have minimal effect on the ILC measurement, except

perhaps in the areas that surround large blood vessels. ILC maps represent a measure of local coherence at each voxel, characterized by the average correlation between each individual voxel and a region of neighbouring voxels. Within the grey matter, ILC has previously been reported⁶² to be found to be higher in the default-mode network.

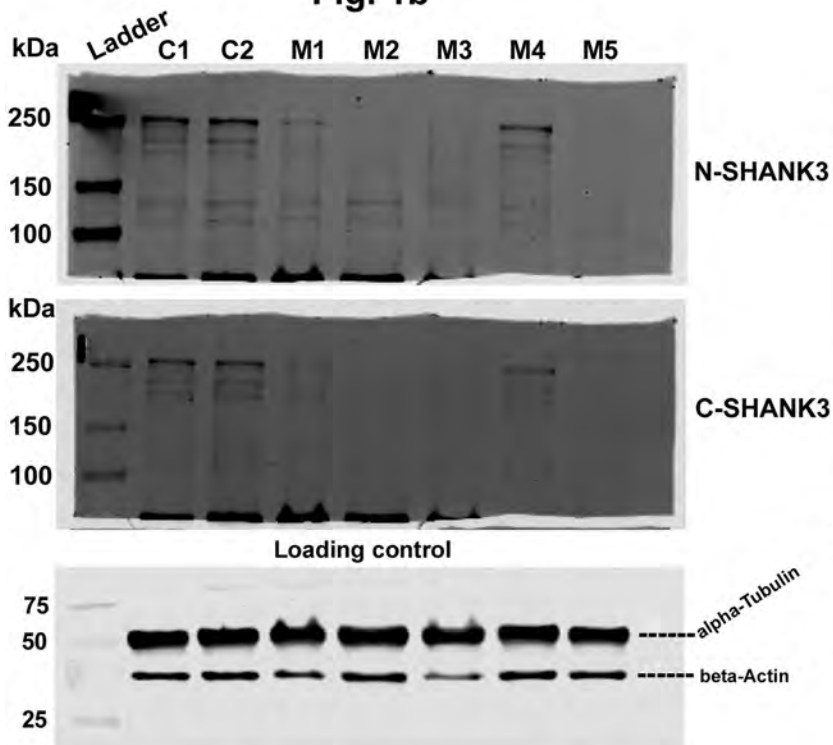
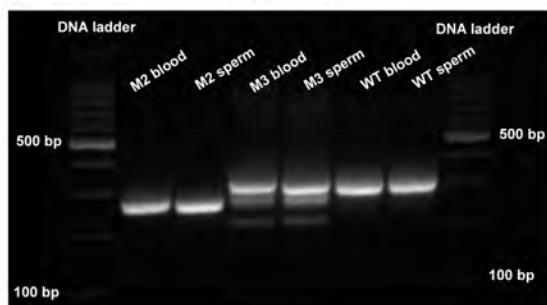
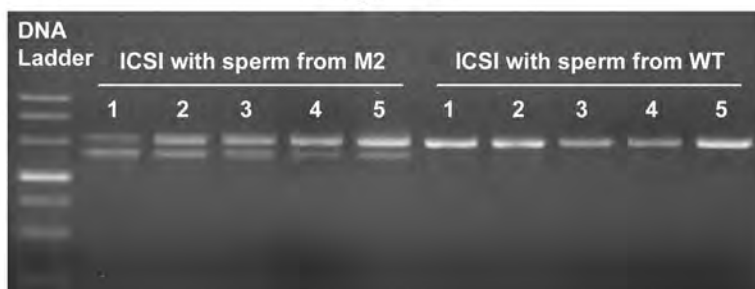
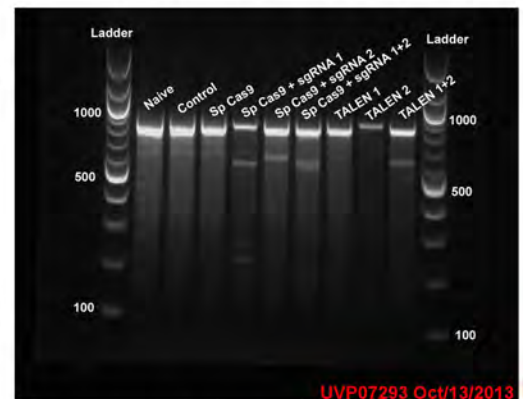
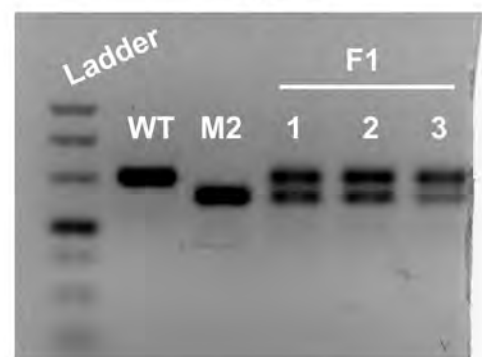
Global correlation. Global correlation (GCOR) is calculated by computing the average of correlation coefficients between each individual voxel and all of the voxels in the brain. GCOR maps provide a voxel-by-voxel measurement of network centrality, characterized by the strength and sign of functional connectivity between a given voxel and every other voxel in the brain. For both ILC and GCOR, the voxel-to-voxel (local and global connectivity) measures are normalized to zero-mean, one-unit variance separately for each subject before being entered into the second-level analyses. Second-level within-group and between-group *t*-tests are performed for the ILC and GCOR maps for both controls and mutants as described in the toolbox for connectivity measures: <https://sites.google.com/view/conn/measures/networks-voxel-level>.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

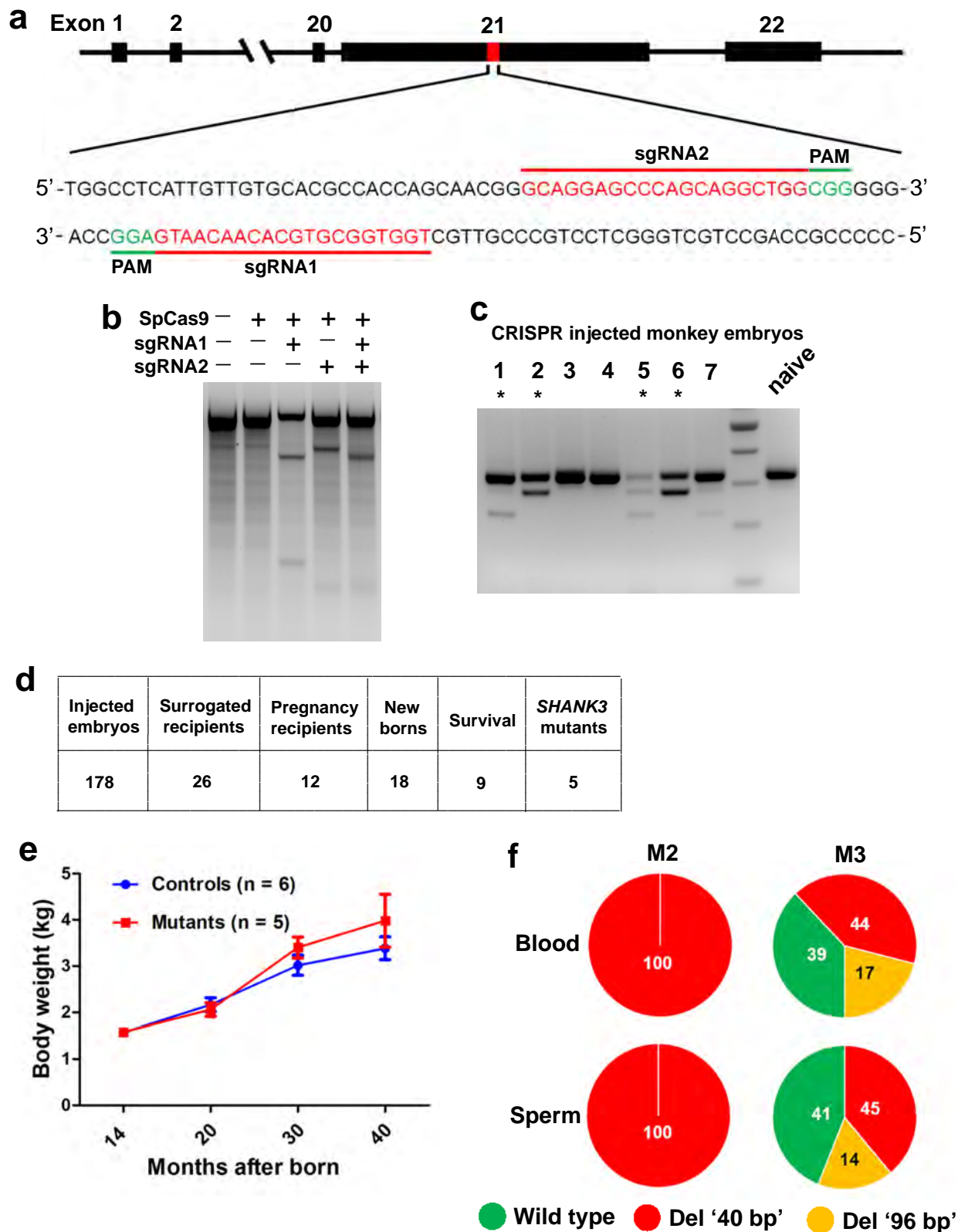
Data availability

All data are available in the main text or the Supplementary Information. All sequencing data, images, code, and materials used in the analysis are available to researchers for the purpose of reproducing or extending the analyses.

51. Ke, Q. et al. TALEN-based generation of a cynomolgus monkey disease model for human microcephaly. *Cell Res.* **26**, 1048–1061 (2016).
52. Sri Kantha, S. & Suzuki, J. Sleep quantitation in common marmoset, cotton top tamarin and squirrel monkey by non-invasive actigraphy. *Comp. Biochem. Physiol. A* **144**, 203–210 (2006).
53. Freund, J. et al. Emergence of individuality in genetically identical mice. *Science* **340**, 756–759 (2013).
54. Bei, D. M. & Lafferty J. D. Dynamic topic models. In *Proc. 23rd International Conference Machine Learning* (2006).
55. Kalman, R. E. A new approach to linear filtering and prediction problems. *J. Basic Engineer.* **82**, 34–45 (1960).
56. Harlow, H. F. & Bromer, J. A. A test apparatus for monkeys. *Psychol. Rec.* **2**, 434–436 (1938).
57. Harlow, H. F. The development of learning in the rhesus monkey. *Am. Sci.* **47**, 459–479 (1959).
58. Levin, E. D. & Bowman, R. E. The effect of pre- or postnatal lead exposure on Hamilton Search Task in monkeys. *Neurobehav. Toxicol. Teratol.* **3**, 391–394 (1983).
59. Frey, S. et al. An MRI based average macaque monkey stereotaxic atlas and space (MNI monkey space). *Neuroimage* **55**, 1435–1442 (2011).
60. Ashburner, J. SPM: a history. *Neuroimage* **62**, 791–800 (2012).
61. Behzadi, Y., Restom, K., Liao, J. & Liu T. T. A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. *Neuroimage* **37**, 90–101 (2007).
62. Deshpande, G., LaConte, S., Peltier, S. & Hu X. Integrated local correlation: a new measure of local coherence in fMRI data. *Hum. Brain Mapp.* **30**, 13–23 (2009).

Fig. 1b**Fig. 1d****Fig. 1e****Extended Data Fig. 2b****Extended Data Fig. 2c****Fig. 1f**

Extended Data Fig. 1 | Original images for western blots and DNA gel. Original images for western blots and DNA gel electrophoresis corresponding to specific figure panels as indicated are presented without cropping or further processing, such as adjusting of brightness and contrast.

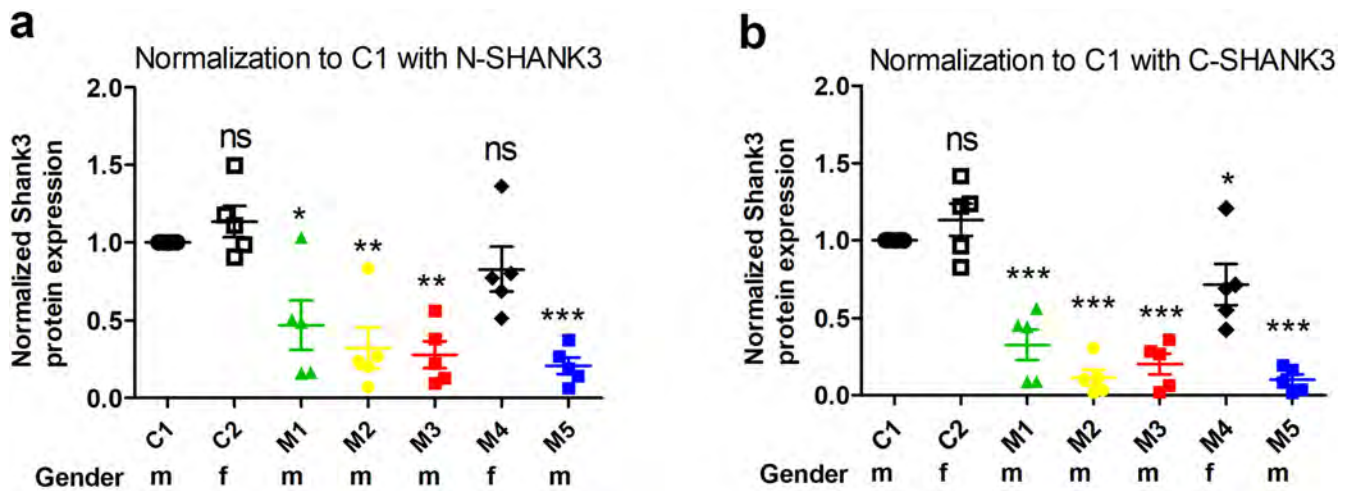


Extended Data Fig. 2 | Summary of founder and germline-transmitted SHANK3 mutations. **a**, Schematic showing the structure of the wild-type macaque SHANK3 gene and magnified panels with the annotated sequence of the gRNA and protospacer adjacent motif (PAM) for both strands within exon 21. **b**, SURVEYOR assay showing SpCas9-mediated indels in cultured cynomolgus monkey primary skin fibroblasts with indicated gRNAs. **c**, Genotyping PCR results of individual monkey

embryos injected with a mixture of SpCas9 mRNA, SHANK3 gRNA no. 1 and gRNA no. 2. Asterisks indicate effectively edited embryos. **d**, Number of injected embryos, transferred recipients and newborn macaques in this study. **e**, SHANK3-mutant macaques have similar body weights to those of age-matched wild-type controls. **f**, Pie charts of genotype (indels) of DNA from semen from mutant macaques M2 and M3 show a similar pattern to their respective blood samples.

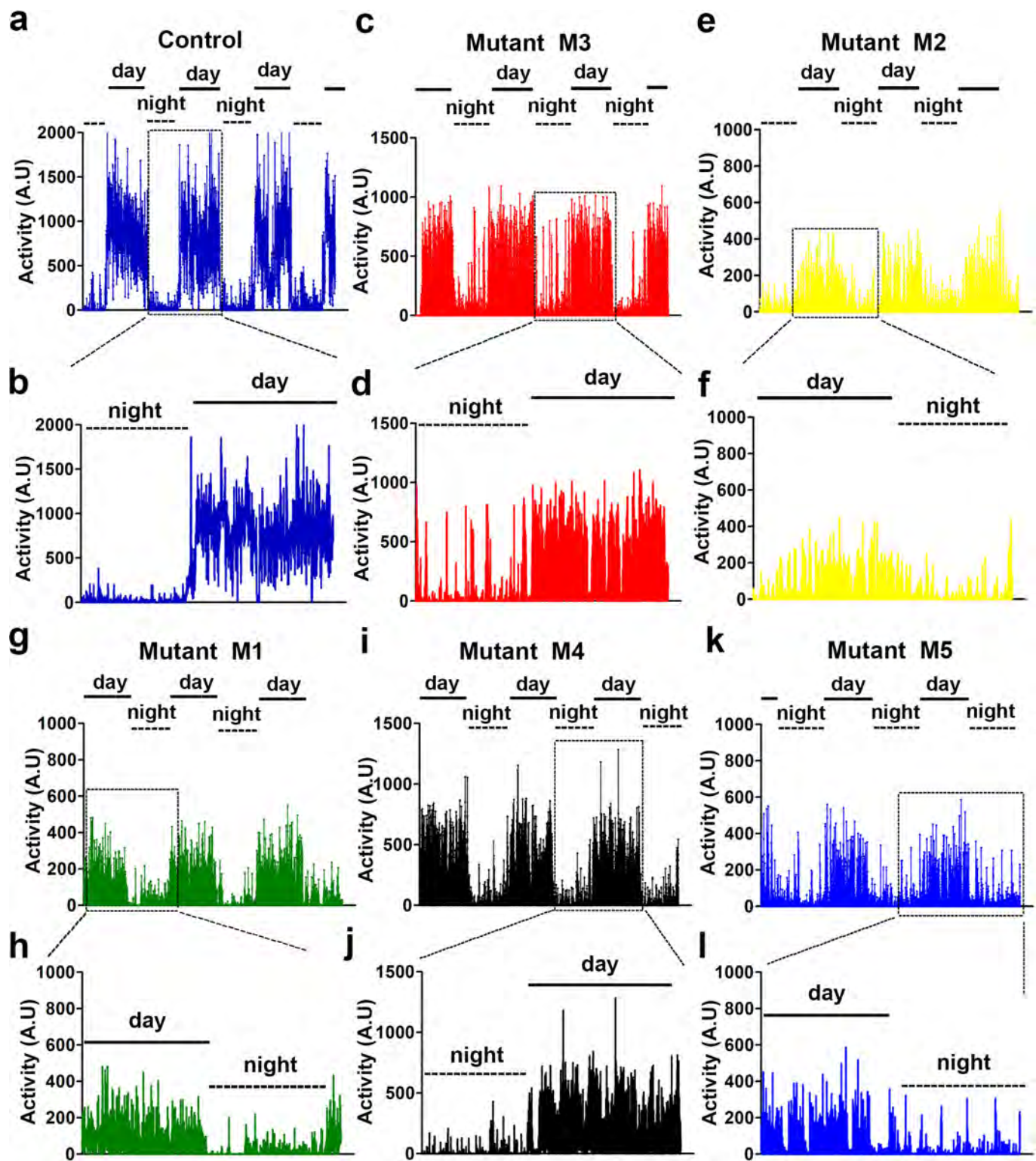
[illegible]

Extended Data Fig. 3 | Alignment of partial *SHANK3* sequence genotyped from skin DNA. a–e, Alignment of ten representative reads of *SHANK3* sequence genotyped from a skin biopsy of each mutant monkey with reference *SHANK3* sequence from wild-type monkey.

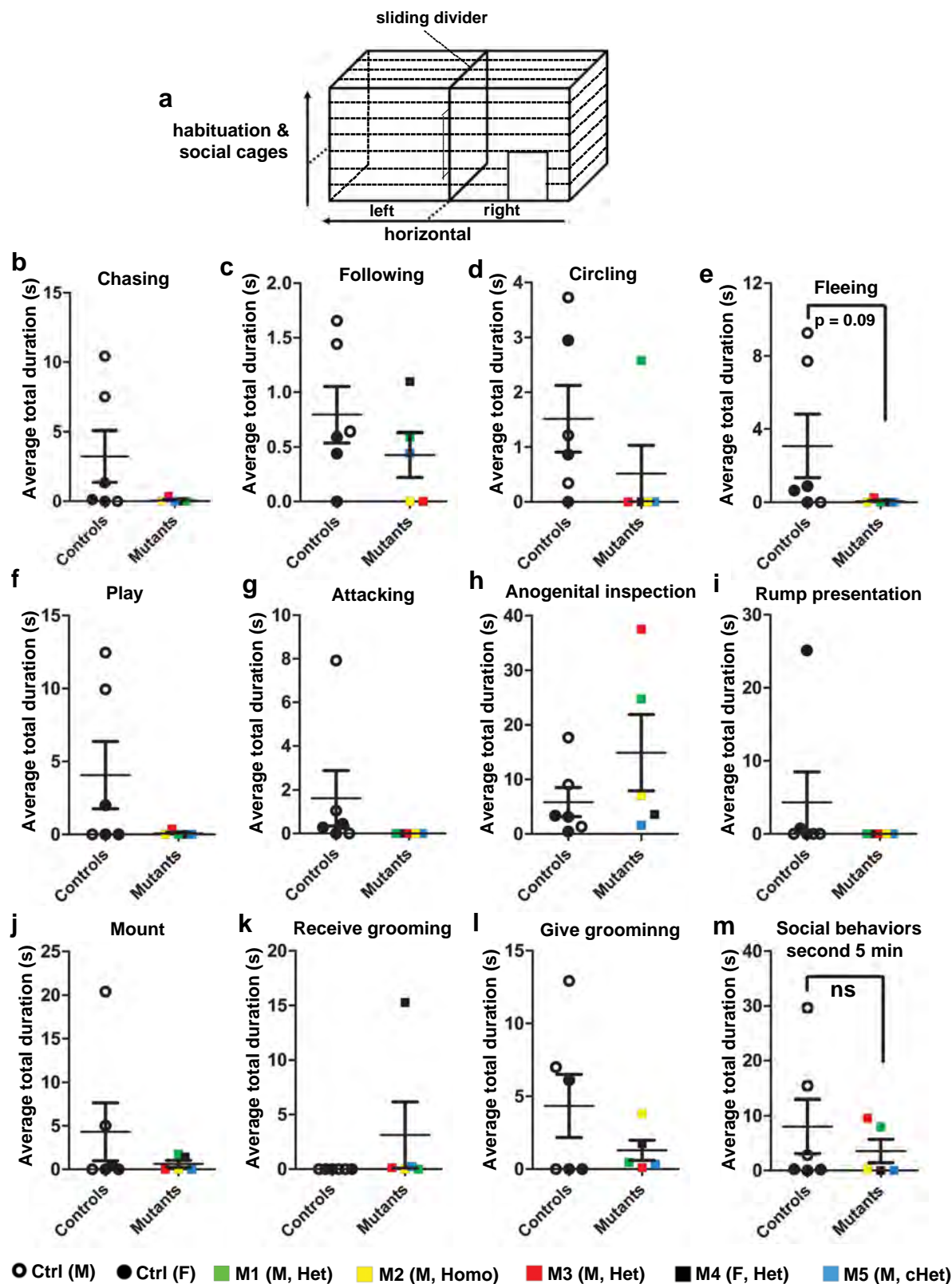


Extended Data Fig. 4 | Statistical analysis of western blots using brain lysates prepared from V1 biopsy of macaques. a, b, Quantification of blots was based on five technical repeats using the same V1 protein sample with N-terminal (a) and C-terminal (b) antibodies. Values were normalized to those of the C1 control monkey. α -Tubulin, as loading

control, was run on the same gel. Data are presented as mean \pm s.e.m., $n = 5$ technical repeats using sample for the 2 controls and 5 *SHANK3* mutants, * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$; ns, not significant; one-way analysis of variance (ANOVA) with Bonferroni post hoc test.

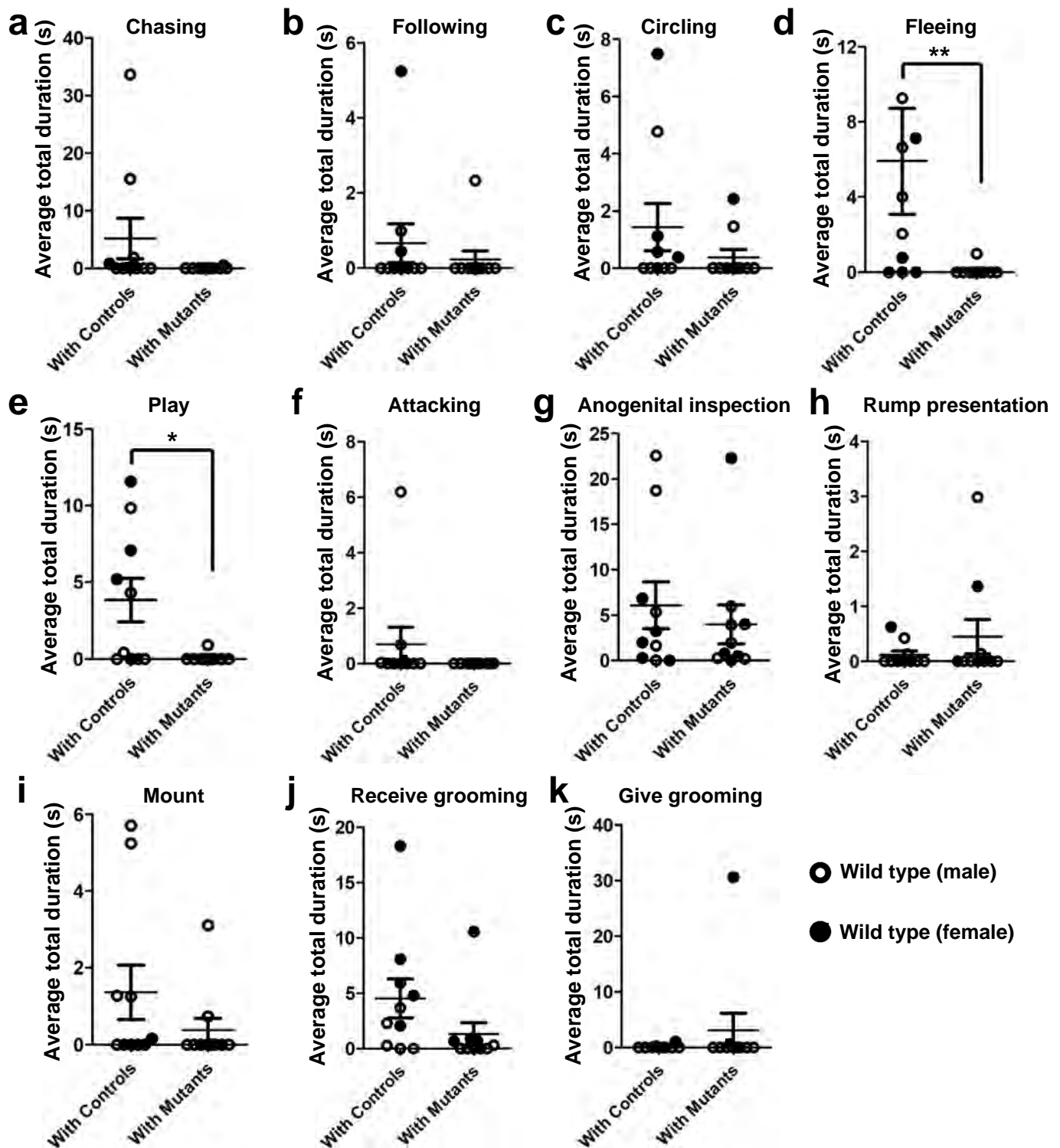


Extended Data Fig. 5 | Representative traces of overall activity. a–l, Representative and enlarged traces of overall activity recorded by motion watches across multiple days from a control macaque and all five SHANK3 mutants. A.U, arbitrary units.



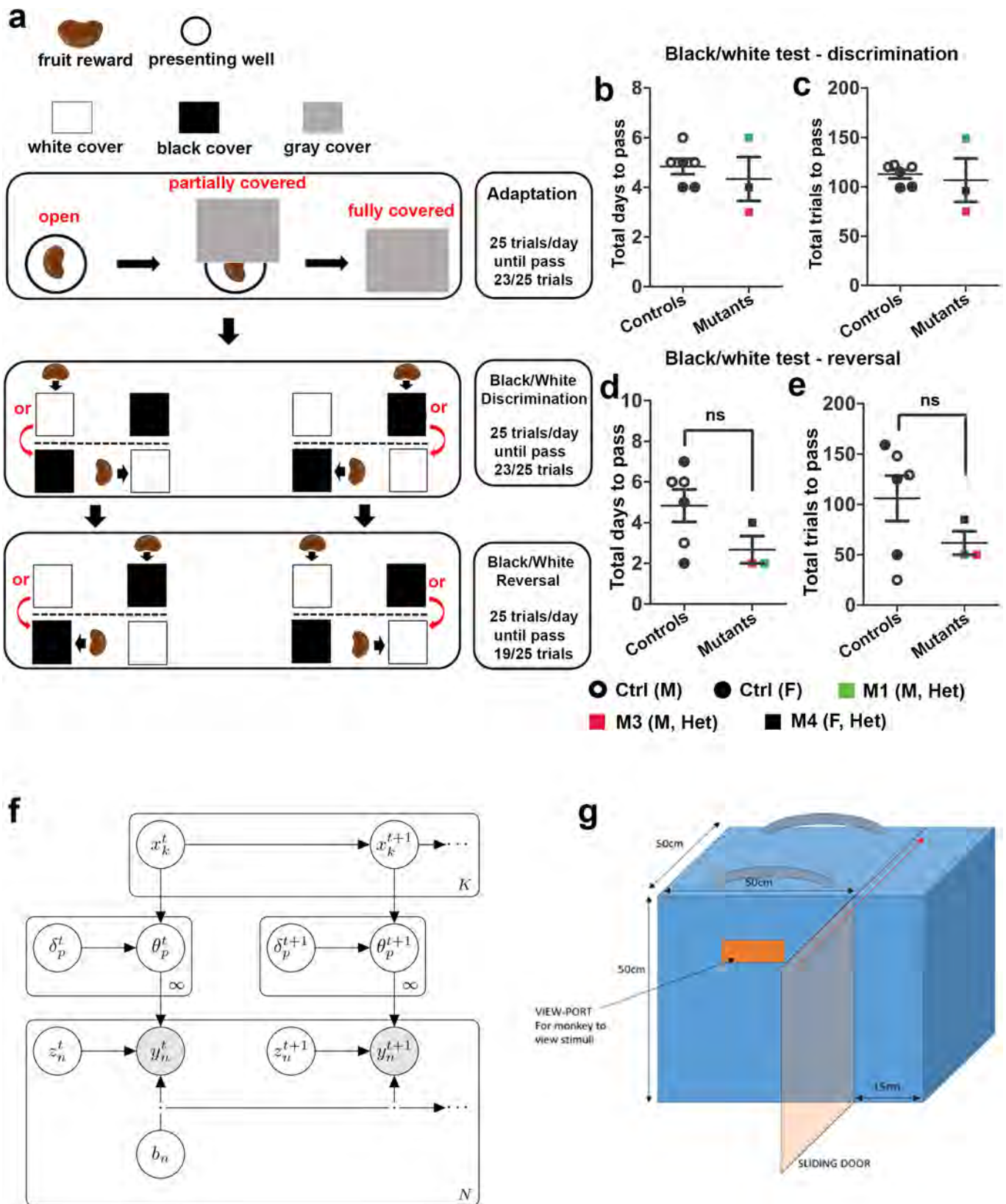
Extended Data Fig. 6 | Behavioural parameters of monkeys during the first and second five minutes of interaction. **a**, Schematic showing the two interconnected cages used for habituation of individual macaques and subsequent paired social-interaction assay. **b–l**, Separate behavioural parameters of monkeys in control and *SHANK3*-mutant groups during

the first five minutes of interaction. **m**, No difference in social behaviours (including chasing, following, circling, fleeing and play) during the second five minutes of interaction. In all panels, $n = 6$ macaques for control group; $n = 5$ macaques for the *SHANK3*-mutant group. Data are presented as mean \pm s.e.m., two-tailed Mann-Whitney *U*-test.



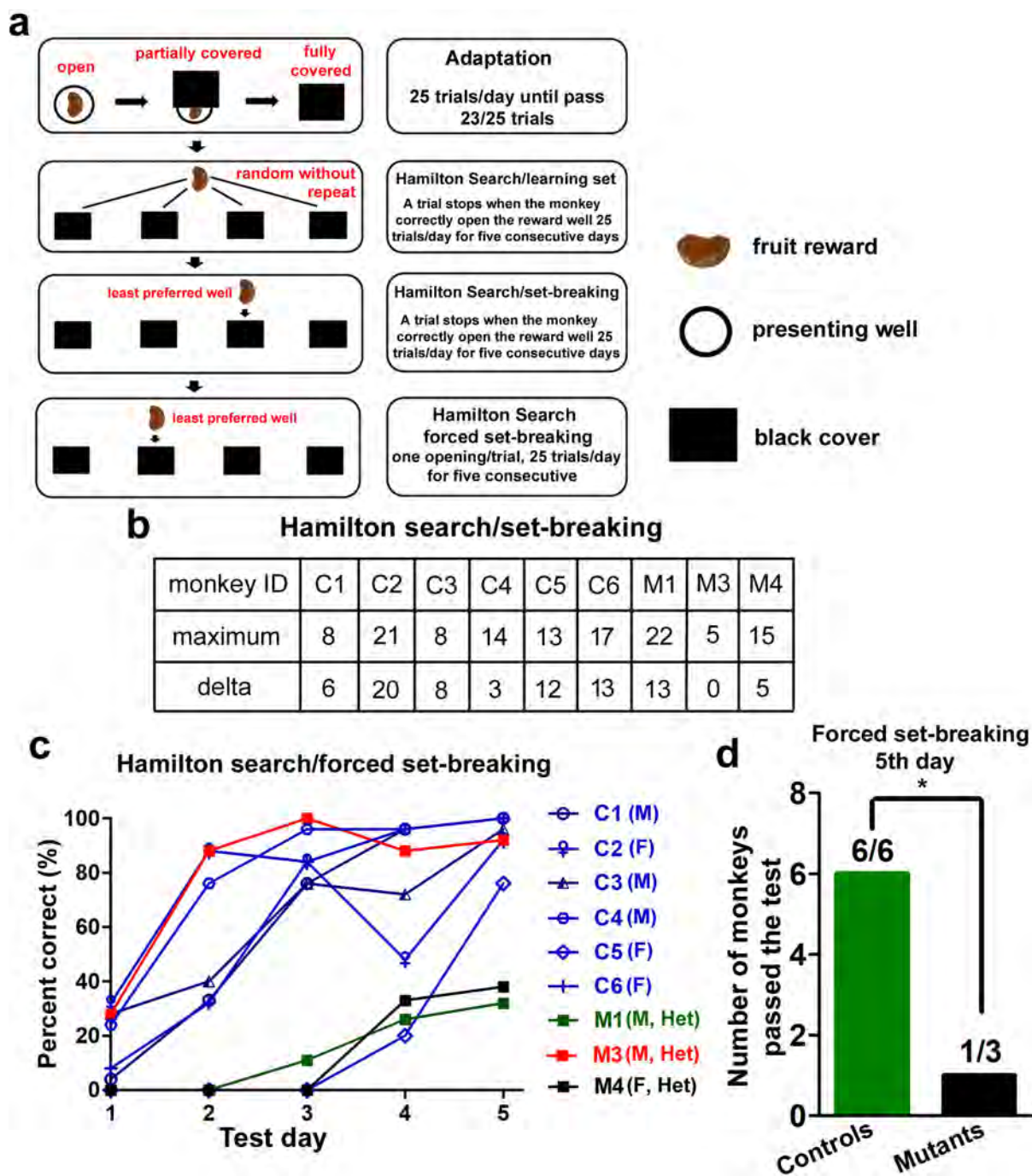
Extended Data Fig. 7 | Behavioural parameters of probe macaques when paired with wild-type or *SHANK3*-mutant monkeys during the first five minutes of interaction. a–k, Total durations of chasing (a), following (b), circling (c), fleeing (d), play (e), attacking (f), anogenital inspection (g), rump presentation (h), mounting (i), receiving grooming (j)

and giving grooming (k). In all panels, $n = 10$ probe monkeys paired individually with 6 wild-type macaques from the control group and 5 macaques from the *SHANK3*-mutant group. Data are presented as mean \pm s.e.m., * $P < 0.05$, ** $P < 0.01$; two-tailed Mann–Whitney U -test.



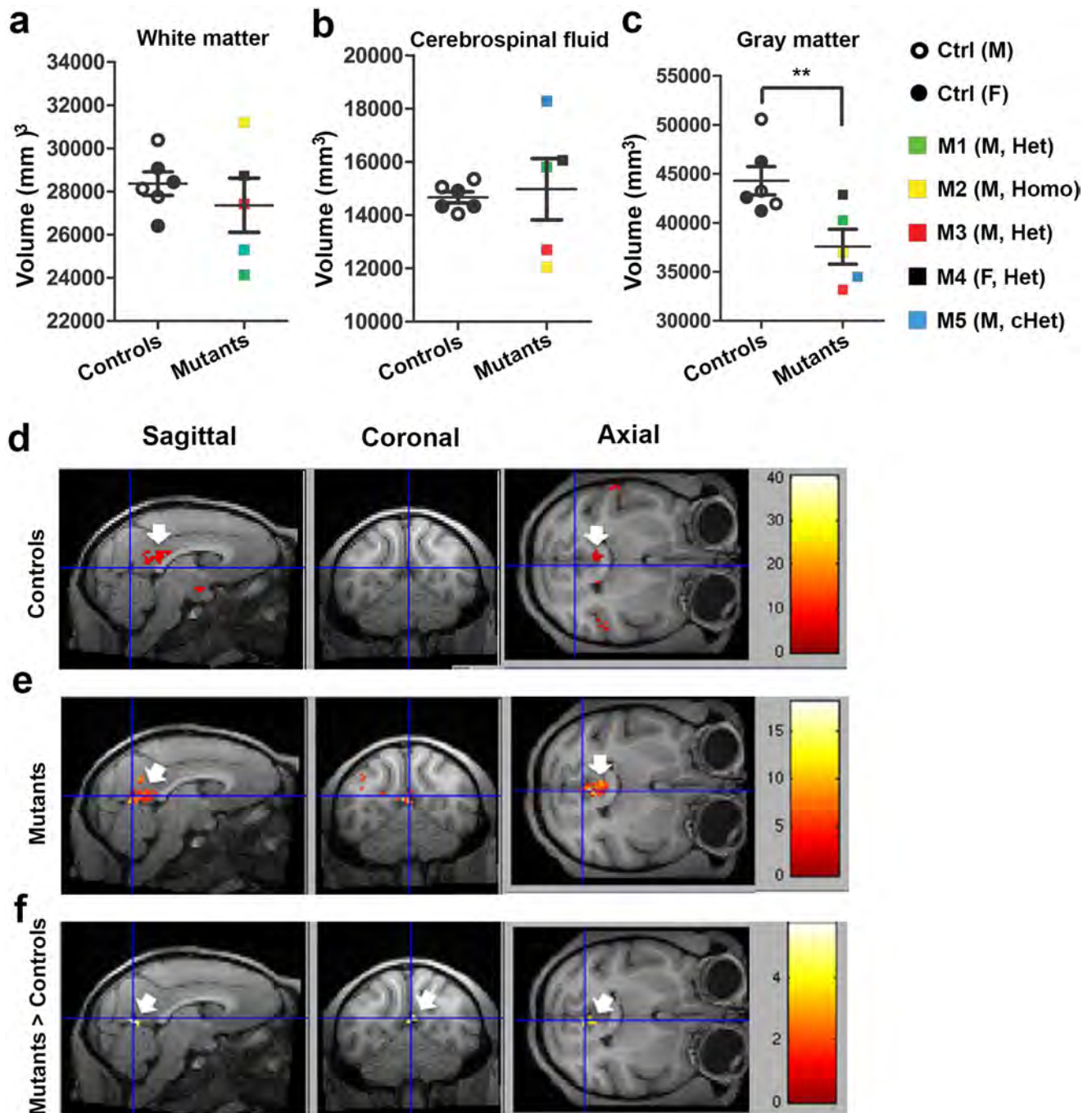
Extended Data Fig. 8 | Performance of control and mutant monkeys in the discrimination and reversal tasks using WGTA. a, Task design. **b, c**, Total days (**b**) and total trials (**c**) required for macaques to pass the black–white discrimination test of the WGTA. **d, e**, Total days (**d**) and total trials (**e**) required for macaques to pass the black–white reversal test of the WGTA (>75%-correct trial). **f**, A graphical model for Bayesian

nonparametric multitarget tracking. Priors omitted for brevity. Arrows pointing to ellipses indicate continuation to the next time step. **g**, Diagram of the eye-tracking box. In **b–e**, $n = 6$ macaques for control group; $n = 3$ macaques for the *SHANK3*-mutant group. Data are presented as mean \pm s.e.m.; Mann–Whitney *U*-test. Coloured squares indicate individual macaques with *SHANK3* mutations.



Extended Data Fig. 9 | Performance of controls and SHANK3 mutants in the Hamilton search task. a, Schematic and flow chart of the Hamilton search task. **b,** Performance of macaques in the ‘set-breaking’ test of the Hamilton search task. M3 showed no improvement (delta value = 0). ‘Delta’ is set to measure the learning of the monkey across five test days, calculated by increase of the number of trials in which the monkey opened

the correct well on the first try. **c,** Percentage of correct trials on the ‘forced set-breaking’ test of the Hamilton search task, from monkeys across five test days. **d,** Number of monkeys that reached a 75%-correct rate on the fifth day of the forced set-breaking test. * $P < 0.05$, Two-tailed χ^2 test ($P = 0.023$) was applied to determine the statistical difference between groups.



Extended Data Fig. 10 | Structural MRI and seed-based functional MRI analysis of macaque brains. **a–c**, No difference in white matter volume (**a**) and cerebrospinal fluid volume (**b**), but a reduced volume of gray matter (**c**), in *SHANK3* mutants, relative to control macaques. In **a–c**, $n = 6$ macaques for control group; $n = 5$ macaques for *SHANK3*-mutant group. Data are presented as mean \pm s.e.m., $**P < 0.01$, Mann–Whitney *U*-test. Coloured squares indicate individual mutant macaques. **d**, **e**, Sagittal,

coronal and axial views of averaged functional MRI image from six control macaques (**d**) and five *SHANK3* mutants (**e**), using the putative posterior cingulate cortex as seed region. **f**, Sagittal, coronal and axial views of averaged functional MRI image show blood-oxygen-level-dependent signals in the posterior cingulate cortex that are greater in mutants than in controls. In **d–f**, the putative posterior cingulate cortex regions are highlighted by arrows.

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a Confirmed

- ☐ ☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
- ☐ ☒ An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- ☐ ☒ The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- ☒ ☐ A description of all covariates tested
- ☒ ☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- ☒ ☐ A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- ☒ ☐ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted
Give P values as exact values whenever suitable.
- ☒ ☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- ☐ ☒ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- ☒ ☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
- ☐ ☒ Clearly defined error bars
State explicitly what error bars represent (e.g. SD, SE, CI)

Our web collection on [statistics for biologists](#) may be useful.

Software and code

Policy information about [availability of computer code](#)

Data collection

Standard programs were used

Data analysis

Prism 5 was used for statistical analysis in this study

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All data are available in the main text or the supplementary materials. All sequencing data, images, code, and materials used in the analysis are available to researchers for the purpose of reproducing or extending the analyses

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/authors/policies/ReportingSummary-flat.pdf](https://www.nature.com/authors/policies/ReportingSummary-flat.pdf)

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No statistical calculations were applied to pre-determine sample size, but our sample sizes are similar to those reported in previous publications (reference # 18: Liu et al., Nature 2016 and reference # 19: Chen et al., Cell 2017).
Data exclusions	No data point from any behavioral or image assays was excluded for all results reported in this study.
Replication	All experimental findings were reliably reproduced among all subjects in all experiments.
Randomization	SHANK3 mutant animals were allocated based on availability after genetic engineering; wild-type control animals were allocated based upon age, and gender similarity to mutants. Randomization of animals and groups were applied throughout all behavioral and MRI assays in the study.
Blinding	All behavioral observations, scoring, data analysis in this study were carried out by trained researchers without prior knowledge of the experimental design and goal of this study. MRI image analysis was relied on objective, automatized measurements, and investigators were blinded to group allocation during data collection and analysis.

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Unique biological materials
<input type="checkbox"/>	<input checked="" type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology
<input type="checkbox"/>	<input checked="" type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Human research participants

Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input type="checkbox"/>	<input checked="" type="checkbox"/> MRI-based neuroimaging

Antibodies

Antibodies used	We utilized N-SHANK3 antibody (1:100 from NIH NeuroMab N367/62), C-SHANK3 antibody (1: 500 from Santa Cruz Biotech SC-30193), and Beta-Tubulin antibody (1: 5000 from Sigma T8328) as primary antibodies. We used goat-anti-mouse IRDye 680 (1: 5000 from Li-COR Biosciences), donkey-anti-rabbit IRDye 800CW (1: 5000 from Li-COR Biosciences) as secondary antibodies.
Validation	All antibodies utilized in this study were previously validated in the reference # 26: Zhou et al. Neuron 2016. Li-COR Odyssey Image Studio software was used to image blots and to quantify the intensity of band automatically. Detailed information of each antibody is listed below: N-SHANK3 antibody (1:100 from NIH NeuroMab N367/62) http://neuromab.ucdavis.edu/datasheet/N367_62.pdf C-SHANK3 antibody (1: 500 from Santa Cruz Biotech SC-30193) http://datasheets.scbt.com/sc-30193.pdf Beta-Tubulin antibody (1: 5000 from Sigma T8328) https://www.sigmaaldrich.com/catalog/product/sigma/t8328?lang=en&region=US Goat-anti-mouse IRDye 680 (1: 5000 from Li-COR Biosciences) https://www.licor.com/documents/802khe2qs9bdc7k88p9Donkey-anti-rabbit-IRDye-800CW-1-5000-from-Li-COR-Biosciences https://www.licor.com/bio/products/reagents/secondary_antibodies/irdye_800cw.html

Animals and other organisms

Policy information about [studies involving animals](#); [ARRIVE guidelines](#) recommended for reporting animal research

Laboratory animals	Juvenile and adult Cynomolgus monkeys (Macaca fascicularis) in both genders were utilized in this study.
--------------------	--

Wild animals

The study did not involve wild animals.

Field-collected samples

The study did not involve samples collected from fields.

Magnetic resonance imaging

Experimental design

Design type

Resting state functional and structural MRI under anesthesia

Design specifications

In all subjects, the slices for functional MRI were acquired using contiguous, interleaved acquisition; 128 volumes per run (3 consecutive runs, each lasting approximately 5 minutes).

Behavioral performance measures

No behavioral measure were taken

Acquisition

Imaging type(s)

Structural and functional

Field strength

3 Tesla

Sequence & imaging parameters

MP-RAGE and T2*-weighted gradient echo, echo planar imaging. Following parameters were used - volume repetition time (TR) = 2.1 s, echo time (TE) = 25ms, flip angle = 90°, 1.25mm x 1.25mm in plane resolution and slice thickness = 1.3mm). T1-weighted, magnetization-prepared rapid gradient echo (MP-RAGE) structural images were also acquired (TR = 2.1s; TE = 3.21ms, flip angle = 8°, 0.5mm isotropic voxels)

Area of acquisition

Whole brain

Diffusion MRI

☐ Used☒ Not used

Preprocessing

Preprocessing software

SPM8, CONN toolbox

Normalization

Data were spatially normalized using segmentation routine in SPM8 to structural scans. Mutual information Affine registration was used for alignment; B-Spline interpolation was used for writing normalized images.

Normalization template

We used MNI Cynmologous macaque atlas as structural templates (purl.org/net/kbmd/cyno).

Noise and artifact removal

Artifact rejection toolbox (ART; www.nitrc.org/projects/artifact_detect) was used for a comprehensive analysis of sources of artifacts in time series data including spiking and motion. Maximum motion (Controls=0.09±0.01 vs. mutants=0.099±0.01; p=0.66); mean motion (controls 0.036 +/- 0.01; mutants 0.036 +/- 0.01; p=0.48). The physiological noise correction was implemented through an anatomical component based noise correction approach (aCompCor) which involves removal of first 3 principal components of the signal from white matter and CSF

Volume censoring

For volume censoring outlier data points (TRs) were defined as volumes that exceeded three z-normalized standard deviations away from mean global brain activation across the entire volume or a composite movement threshold of 0.5 mm scan-to-scan frame-wise displacement.

Statistical modeling & inference

Model type and settings

For resting state local and global functional connectivity analysis, we employed an unbiased data driven approach. Integrated Local Correlation (ILC) was used to assess coupling of local neuronal processes in any given voxel's neighborhood. ILC maps represent a measure of local coherence at each voxel, characterized by the average correlation between each individual voxel and a region of neighboring voxels.

For both ILC and GCOR, first-level correlation maps were produced by extracting denoised time series by regressing non-BOLD components from each voxel to the rest of the voxels, followed by computing Pearson's correlation coefficients (r) between that time course and the time course of all other voxels. Correlation coefficients were converted to normally distributed z-scores using Fisher's r-to-z transformation to perform second-level General Linear Model (GLM) analyses. These second-level within group and between-group t-tests were performed for the correlation maps for both controls and mutants (<https://sites.google.com/view/conn/measures/networks-voxel-level>).

Effect(s) tested

There were no tasks or stimulus conditions. Global Correlation (GCOR) and Local correlations (ILC) were used for resting state functional connectivity analysis. GCOR maps provide a voxel by voxel measure of network centrality, characterized by the strength and sign of functional connectivity between a given voxel and every other voxel in the brain. ILC maps represent a measure of local coherence at each voxel, characterized by the average correlation between each individual voxel and a region of neighboring voxels.

Specify type of analysis: ☒ Whole brain ☐ ROI-based ☐ BothStatistic type for inference
(See [Eklund et al. 2016](#))

Voxel to voxel analysis.

Correction

Height threshold = $p < 0.005$ and cluster threshold = $p < 0.05$ (corrected for false discovery rate)

Models & analysis

n/a | Involved in the study

- ☐ ☒ Functional and/or effective connectivity
- ☒ ☐ Graph analysis
- ☒ ☐ Multivariate modeling or predictive analysis

Functional and/or effective connectivity

For resting state local and global functional connectivity analysis, we employed an unbiased data driven approach. Integrated Local Correlation (ILC) was used to assess coupling of local neuronal processes in any given voxel's neighborhood. ILC maps represent a measure of local coherence at each voxel, characterized by the average correlation between each individual voxel and a region of neighboring voxels.

For both ILC and GCOR, first-level correlation maps were produced by extracting denoised time series by regressing non-BOLD components from each voxel to the rest of the voxels, followed by computing Pearson's correlation coefficients (r) between that time course and the time course of all other voxels. Correlation coefficients were converted to normally distributed z-scores using Fisher's r-to-z transformation to perform second-level General Linear Model (GLM) analyses. These second-level within group and between-group t-tests were performed for the correlation maps for both controls and mutants (<https://sites.google.com/view/conn/measures/networks-voxel-level>).

Single-cell transcriptomic analysis of Alzheimer's disease

Hansruedi Mathys^{1,2,10}, Jose Davila-Velderrain^{3,4,10}, Zhuyu Peng^{1,2}, Fan Gao^{1,2}, Shahin Mohammadi^{3,4}, Jennie Z. Young^{1,2}, Madhvi Menon^{4,5,6}, Liang He^{3,4}, Fatema Abdurrob^{1,2}, Xueqiao Jiang^{1,2}, Anthony J. Martorell^{1,2}, Richard M. Ransohoff⁷, Brian P. Hafler^{4,5,6,8}, David A. Bennett⁹, Manolis Kellis^{3,4,11*} & Li-Huei Tsai^{1,2,4,11*}

Alzheimer's disease is a pervasive neurodegenerative disorder, the molecular complexity of which remains poorly understood. Here, we analysed 80,660 single-nucleus transcriptomes from the prefrontal cortex of 48 individuals with varying degrees of Alzheimer's disease pathology. Across six major brain cell types, we identified transcriptionally distinct subpopulations, including those associated with pathology and characterized by regulators of myelination, inflammation, and neuron survival. The strongest disease-associated changes appeared early in pathological progression and were highly cell-type specific, whereas genes upregulated at late stages were common across cell types and primarily involved in the global stress response. Notably, we found that female cells were overrepresented in disease-associated subpopulations, and that transcriptional responses were substantially different between sexes in several cell types, including oligodendrocytes. Overall, myelination-related processes were recurrently perturbed in multiple cell types, suggesting that myelination has a key role in Alzheimer's disease pathophysiology. Our single-cell transcriptomic resource provides a blueprint for interrogating the molecular and cellular basis of Alzheimer's disease.

Alzheimer's disease (AD) is a slowly progressing neurodegenerative disorder that starts with mild memory loss and culminates in severe impairment of executive and cognitive functions^{1–3}. The pathophysiology of AD involves interactions between neuron and glia; in support of this, transcriptomic and epigenomic analyses reveal downregulation of neuronal functions and upregulation of innate immune responses in AD brains^{4–14}. However, bulk-tissue-level resolution may mask the complexity of alterations across cells and within cell groups, especially for less abundant cell types⁴. Potential changes in cell composition during neurodegeneration also confound the distinction between composition and activity changes in a given cell type. Moreover, the complex interplay between protective and damaging molecular processes, both within and across cell types, further contributes to the difficulty in interpreting tissue-resolution signatures of disease.

Single-cell RNA sequencing (scRNA-seq) provides an alternative method to study the cellular heterogeneity of the brain^{15–17}, by profiling tens of thousands of individual cells^{15,18,19}. With the goal of characterizing the complex cellular changes in AD brain pathology, here we profile 80,660 droplet-based single-nucleus cortical transcriptomes across 48 individuals with varying degrees of AD pathology and including both sexes. The resulting resource—which is the first, to our knowledge, single-cell view of AD pathology—paints a unique cellular-level view of transcriptional alterations associated with AD pathology, and reveals cell-type-specific and shared gene-expression perturbations, disease-associated cellular subpopulations, and sex-biased transcriptional responses.

Single-nucleus RNA-seq profiling of prefrontal cortex

Post-mortem human brain samples came from 48 participants in the Religious Order Study (ROS) or the Rush Memory and Aging Project (MAP), two longitudinal cohort studies of ageing and dementia.

Information collected as part of these studies (which are collectively known as ROSMAP) includes clinical data, detailed post-mortem pathological evaluations, and omics tissue profiling of participants²⁰. We selected 24 individuals with high levels of β -amyloid and other pathological hallmarks of AD ('AD-pathology'), and 24 individuals with no or very low β -amyloid burden or other pathologies ('no-pathology'). For each individual, we profiled tissue from the prefrontal cortex (Brodmann area 10)—a region of the brain that has a major role in traits that are affected by AD, including cognition. Immunohistochemistry for β -amyloid confirmed the pathological status of the samples (Extended Data Fig. 1a, b), and bright-field and high-resolution confocal microscopy did not show any apparent physical damage to nuclei isolated from AD-pathology samples relative to no-pathology samples (Extended Data Fig. 1c). We report a total of 80,660 droplet-based single-nucleus RNA-seq (snRNA-seq) profiles (Extended Data Fig. 2a), which are publicly available on the ROSMAP data compendium (see 'Data availability').

Cellular diversity of the aged prefrontal cortex

To classify the major cell types in the aged cortex, we pre-clustered all cells jointly across the 48 individuals (Extended Data Fig. 2b) to produce 20 transcriptionally distinct pre-clusters with highly consistent expression patterns across individuals (Extended Data Fig. 2c, d). We identified and annotated the major cell types of the human brain by interrogating the expression patterns of known marker genes^{18,21}: excitatory neurons (marked by *NRGN*), inhibitory neurons (*GAD1*), astrocytes (*AQP4*), oligodendrocytes (*MBP*), microglia (*CSF1R* and *CD74*), oligodendrocyte progenitor cells (*VCAN*), endothelial cells (*FLT1*), and pericytes (*AMBP*) (Extended Data Fig. 3a, b). The cell types, markers, and proportions of cell types matched previous single-nucleus droplet-based sequencing data from adult human cortex¹⁸,

¹Picower Institute for Learning and Memory, Massachusetts Institute of Technology, Cambridge, MA, USA. ²Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA. ³MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, MA, USA. ⁴Broad Institute of MIT and Harvard, Cambridge, MA, USA. ⁵Department of Neurology, Harvard Medical School, Boston, MA, USA. ⁶Evergrande Center for Immunologic Diseases, Harvard Medical School, Boston, MA, USA. ⁷Third Rock Ventures, Boston, MA, USA. ⁸Department of Ophthalmology, Harvard Medical School, Boston, MA, USA. ⁹Rush Alzheimer's Disease Center, Rush University Medical Center, Chicago, IL, USA. ¹⁰These authors contributed equally: Hansruedi Mathys, Jose Davila-Velderrain. ¹¹These authors jointly supervised this work: Manolis Kellis, Li-Huei Tsai. *e-mail: manoli@mit.edu; lhtsai@mit.edu

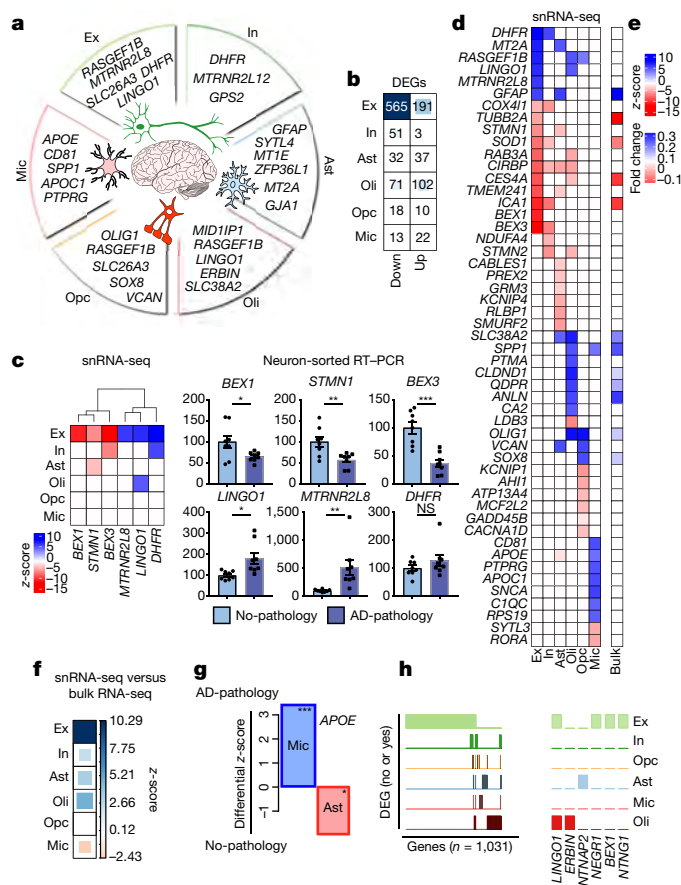


Fig. 1 | Cell-type-specific gene-expression changes in AD pathology. **a**, Genes most upregulated: excitatory (Ex) and inhibitory (In) neurons, astrocytes (Ast), oligodendrocytes (Oli), oligodendrocyte precursor cells (Opc), and microglia (Mic). **b**, DEG counts for each cell type (two-sided Wilcoxon rank-sum test, FDR < 0.01, log₂(mean gene expression in AD-pathology/mean gene expression in no-pathology) > 0.25, Poisson mixed-model FDR < 0.05). The intensity of the blue colour and the size of the squares are proportional to entry values. **c**, RT-qPCR validation. snRNA-seq differential scores for excitatory- and inhibitory-neuron DEGs (z-score, Poisson mixed model) (left) and qPCR validation (right). NeuN-positive nuclei were isolated by fluorescence-activated cell sorting (n = 8 AD-pathology individuals and n = 8 no-pathology individuals). Data are mean ± s.e.m.; ***P < 0.001, **P < 0.01, *P < 0.05; NS, P > 0.05 (Student's two-tailed t-test). **d, e**, Differential scores (z-scores, Poisson mixed model) for the top DEGs from snRNA-seq (d), and the corresponding values from bulk RNA-seq (e) (ROSMAP cohorts; n = 484, P < 0.01). **f**, Global consistency analysis of snRNA-seq versus bulk RNA-seq. Agreement or disagreement was estimated by deviation from random expectation (z-score) of single-cell DEG average rank scores in the ranked list of genes from bulk differential analysis. **g**, Differential expression of *APOE* in microglia (n = 955 cells, 24 AD-pathology individuals; n = 965 cells, 24 no-pathology individuals) and astrocytes (n = 1,830 cells, 24 AD-pathology individuals; n = 1,562 cells, 24 no-pathology individuals; z-score, Poisson mixed model; ***P = 0.001, *P = 0.05, corresponding P value calculated from two-sided, standard normal distribution). **h**, Binary plots indicating with bars whether a gene (column) is a DEG in a given cell type (rows) or not (n = 1,031 DEGs). Right, six genes associated with myelination and/or axon regeneration.

indicating that our results are robust to the inclusion of pathologically affected brains (Extended Data Fig. 3c–e). We next collapsed the pre-clusters into eight broad cell-type clusters using annotations supported by both direct marker expression and significant (false discovery rate (FDR) < 0.01, hypergeometric test) overlap with previously curated single-cell populations. We used these cell-type categories to characterize the specificity of AD-pathology gene-expression perturbations, to quantify gene–trait associations, and to assess

qualitative differences in cell-type-specific pathological responses between sexes.

Systematic differential analysis of gene expression

We compared levels of gene expression in cells isolated from AD-pathology versus no-pathology individuals by cell type (Methods), and identified 1,031 unique differentially expressed genes (DEGs) that implicated all major cell types (Fig. 1a, b, Supplementary Table 2). Neurons showed a strong signature of repression—75% of DEGs in excitatory and 95% in inhibitory neurons were downregulated—whereas most DEGs in oligodendrocytes, astrocytes, and microglia were upregulated (53–63%). The numbers of DEGs for non-neuronal populations were substantially smaller, probably owing to reduced power in lower-abundance cell types. These contrasting observations on the number and dominant directionality of DEGs reveal a heterogeneous response to AD pathology between cell types—a recurrent theme that we observed throughout the study. Notably, DEGs were robustly detected at different levels of expression (Extended Data Fig. 4a). Quantitative PCR with reverse transcription (RT-qPCR) in nuclei staining positive for the neuronal marker NeuN (isolated by fluorescence-activated cell sorting) corroborated significant differential expression for five of the six genes tested, including both upregulated and downregulated genes (Fig. 1c).

A similar analysis of DEGs using high-quality bulk RNA sequencing (RNA-seq) data from the ROSMAP cohorts (n = 484) validated the DEGs that we identified in our snRNA-seq data (Fig. 1d, e). In addition, a rank permutation test (Methods) revealed that bulk data are dominated by expression changes observed at the single-cell level in excitatory neurons and oligodendrocytes (Fig. 1f). Consequently, changes in other cell types are not well captured, with microglia particularly underrepresented. Bulk RNA-seq data also cannot capture DEGs with opposite directionality in different cell types. With snRNA-seq, we found that *APOE* was strongly upregulated in microglia but downregulated in astrocytes (FDR-corrected P < 0.01, two-sided Wilcoxon rank-sum test) (Fig. 1g), which is consistent with studies of AD in human induced pluripotent stem cells²² and in mouse models^{23–25}.

The vast majority of DEGs (95%) were perturbed only in neurons or in a single glial cell type, which indicates that these perturbations are strongly cell-type specific (Fig. 1h). However, we found that top DEGs are involved in related processes even across cell types. One such example is myelination, axonal outgrowth and regeneration. Top DEGs included *LINGO1*, which was upregulated in excitatory neurons and oligodendrocytes and is a negative regulator of neuronal survival, axonal integrity, and oligodendrocyte differentiation and myelination^{26,27}; *ERBIN*, which is required for remyelination of axons²⁸; *CNTNAP2*, which mediates axon–myelin interactions²⁹; *NEGR1*, which mediates myelin-stimulated outgrowth of axons³⁰; *BEX1*, which is involved in regeneration of axons after injury³¹; and *NTNG1*, which promotes neurite outgrowth of axons and dendrites³² (Fig. 1h). Consistent with the observed downregulation of *NTNG1* in excitatory neurons, RNA in situ hybridization revealed significantly fewer excitatory neurons (marked by expression of *SLC17A7*) with detectable *NTNG1* expression in brain sections from AD-pathology individuals compared with no-pathology individuals (Extended Data Fig. 4b).

Overall, these results indicate that all major cell types are affected at the transcriptional level by AD pathology, and that single-cell-level resolution is critical because changes in gene expression—including directionality—can be conditional on cell type.

Cell-type-specific changes during AD progression

We next examined whether there are qualitative changes in gene-expression perturbations during early- versus late-stage AD pathology. To define pathology groups, we clustered individuals based on nine clinico-pathological traits (Supplementary Table 3). AD-pathology individuals segregated into two subgroups that correspond to the pathological progression of AD: ‘early-pathology’¹ (amyloid burden, but modest neurofibrillary tangles and modest cognitive impairment)

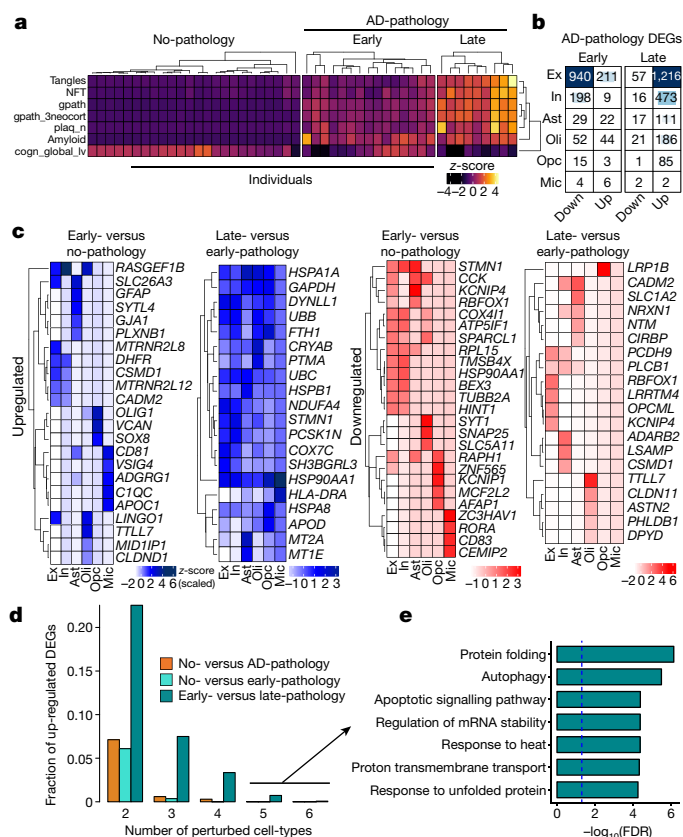


Fig. 2 | Gene-expression changes in the progression of AD pathology.

a, Phenotypic clustering of 48 individuals (columns) using clinico-pathological variables (rows) measuring neuronal neurofibrillary tangle density (tangles), neurofibrillary tangle burden (NFT), global AD-pathology burden (gpath), global measure of neocortical pathology (gpath_3neocort), neuritic plaque burden (plaq_n), overall amyloid level (amyloid), and global cognitive function (last valid score) (cogn_global_iv). **b**, Progressive changes in DEG counts for each cell type for early and late AD-pathology individuals (two-sided Wilcoxon rank-sum test, $FDR < 0.01$, $\log_2(\text{mean gene expression in early-pathology/mean gene expression in no-pathology}) > 0.25$ or $\log_2(\text{mean gene expression in late-pathology/mean gene expression in early-pathology})$ and similarly for the opposite direction, Poisson mixed-model $FDR < 0.05$). The intensity of the blue colour and the size of the squares are proportional to entry values. **c**, Most-significantly altered (based on *P* value rank) genes (rows) for each cell type (columns) and comparison. Column-scaled z-scores computed with a Poisson mixed model are shown ($FDR < 0.01$, two-sided Wilcoxon rank-sum test). **d**, Fraction of total upregulated genes (y axis) as a function of the total number of cell types in which the upregulation occurs. **e**, Gene Ontology terms associated with genes upregulated in the late-pathology group that are common to ≥ 5 cell types ($n = 11$ genes, hypergeometric test, FDR correction).

and 'late-pathology'³³ (higher amyloid burden, and also increased neurofibrillary tangles, global pathology, and cognitive impairment) (Fig. 2a). We then quantified gene-expression changes by pairwise comparisons between these groups.

Comparison of the early-pathology and no-pathology subgroups revealed that large-scale transcriptional changes occur before individuals develop severe pathological features (Fig. 2b, c). Both up- and downregulated DEGs were highly cell-type specific, with nearly all genes (96%) perturbed either in neurons (excitatory and inhibitory) or in a single glial cell type. These changes were similar to those found between the no-pathology and AD-pathology groups (Extended Data Fig. 5), suggesting that major transcriptional changes appear early in pathological progression. Comparison of the late-pathology and early-pathology groups revealed common upregulated genes that were shared across cell types (Fig. 2d), in contrast with the cell-type

specificity that we observed at an early stage of pathology. Shared upregulated genes are involved in protein folding (for example, *HSP90AA1* and *HSPA1A*), including molecular chaperones (for example, *HSPB1* and *CRYAB*), and are also associated with autophagy, apoptosis, and the generalized stress response (Fig. 2c, e). These processes are collectively involved in the proteostasis network—the molecular machinery that operates to maintain protein integrity^{34–36}. In contrast with upregulated genes, downregulated genes mostly showed cell-type-specific changes.

Cell-type-specific associations with AD-related traits

Given the complexity and heterogeneity of the AD phenotype, we next aimed to quantify the association between gene expression in specific cell types and variability of pathological traits. We used the major pathological quantitative traits from the ROS: β -amyloid level, neurofibrillary tangle burden, neuritic plaque count, tangle density, global AD pathology, and global cognition. For each individual, we first computed the correlation between the expression profile of each gene (for a given cell type) and a pathological trait. We then analysed the resultant gene–trait correlation patterns using the self-organizing map (SOM) approach³⁷ to discover gene sets with similar expression patterns that most strongly correlated with each phenotype (Methods). Genes with similar phenotypic correlations are grouped in the same SOM grid unit, with similar units clustered nearby (Extended Data Fig. 6a–e). We observed that excitatory neurons, inhibitory neurons, astrocytes, microglia, and oligodendrocytes each showed distinct SOM units associated with multiple pathological traits, which indicates that different groups of genes respond to AD pathology in each cell type. Gene sets that correlated with post-mortem interval and age at death were highly distinct from those that correlated with pathological signatures of AD, suggesting that they are orthogonal to each other.

To identify gene groups that showed similar correlations with traits, we used an image segmentation method to identify and manually curate the ten SOM territories that were most strongly associated with AD phenotypes across cell types (Extended Data Fig. 6f). We found that the gene groups defining these territories (gene–trait correlation modules M1–M10) are involved in common functional pathways that are relevant to the pathophysiology of AD (Supplementary Table 4). For example, M7 showed a positive pathological correlation in microglia and was enriched in immune and inflammatory pathways, as well as in pathways associated with the clearance of β -amyloid. M9 was positively correlated with pathology in oligodendrocytes and was enriched in oligodendrocyte differentiation and myelination pathways—possibly reflecting an oligodendrocytic response to myelin loss.

To link genetic risk with coordinated gene activity, we quantified the overrepresentation of genes from genome-wide association studies (GWAS)³⁸ and identified modules that were linked to genetic risk factors for AD and for general cognitive function. M6 and M7 overlapped with genes that are AD risk factors, including *APOE*, *TREM2*, *MEF2C*, *PICALM*, and the major histocompatibility complex (MHC) class II genes *HLA-DRB1* and *HLA-DRB5*; expression of these genes in microglia is positively correlated with measures of AD pathology. By contrast, M3 overlapped with genes that are associated with cognition³⁹ and the expression of which is negatively correlated with pathology in neurons (Extended Data Fig. 6g, Supplementary Table 4). These observations provide a link between genetic risk factors and the microglial and neuronal transcriptional responses to pathology, which may partially explain some of the risk conferred by genetic variants.

AD-associated cellular subpopulations

To dissect cell-type heterogeneity, we next sub-clustered each major cell type⁴⁰, resulting in 13 excitatory-neuron (Ex), 12 inhibitory-neuron (In), 4 astrocyte (Ast), 5 oligodendrocyte (Oli), 3 oligodendrocyte-precursor-cell (Opc), and 4 microglia (Mic) sub-clusters. The identified subpopulations were not exclusively enriched with cells from any single individual (Extended Data Fig. 7a, b). We examined the cellular composition of each subpopulation in relation to the pathological features

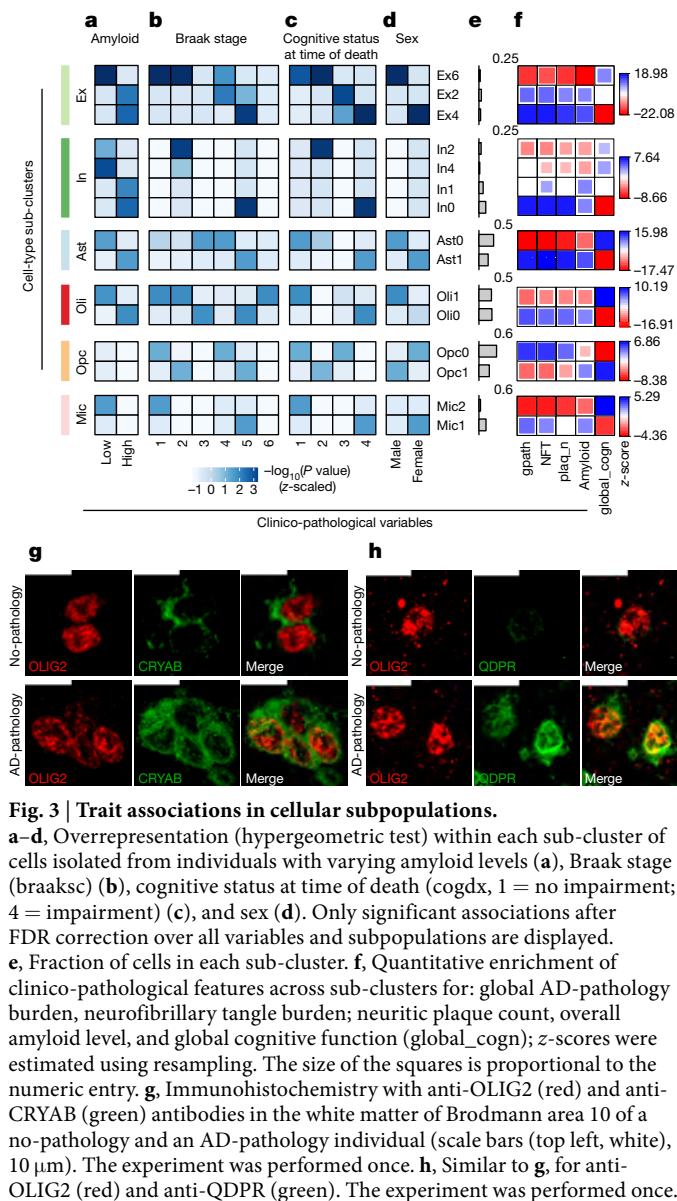


Fig. 3 | Trait associations in cellular subpopulations.

a–d, Overrepresentation (hypergeometric test) within each sub-cluster of cells isolated from individuals with varying amyloid levels (**a**), Braak stage (braaksc) (**b**), cognitive status at time of death (cogdx, 1 = no impairment; 4 = impairment) (**c**), and sex (**d**). Only significant associations after FDR correction over all variables and subpopulations are displayed. **e**, Fraction of cells in each sub-cluster. **f**, Quantitative enrichment of clinico-pathological features across sub-clusters for: global AD-pathology burden, neurofibrillary tangle burden; neuritic plaque count, overall amyloid level, and global cognitive function (global_cogn); z-scores were estimated using resampling. The size of the squares is proportional to the numeric entry. **g**, Immunohistochemistry with anti-OLIG2 (red) and anti-CRYAB (green) antibodies in the white matter of Brodmann area 10 of a no-pathology and an AD-pathology individual (scale bars (top left, white), 10 μm). The experiment was performed once. **h**, Similar to **g**, for anti-OLIG2 (red) and anti-QDPR (green). The experiment was performed once.

of source brains (Methods), and observed an overrepresentation of cellular subtypes in pathology for most major cell types. For example, the cellular subpopulations Ex4, In0, Ast1, and Oli0 were associated with cells that were isolated from subjects with AD-pathology traits—that is, high amyloid level, high Braak stage (V), low CERAD (Consortium to Establish a Registry for Alzheimer's Disease) score, low NIA (National Institute on Aging)–Reagan score, and pronounced cognitive decline. By contrast, the subpopulations Ex6, In2, Ast0, and Oli1 were associated with cells from subjects with no pathological traits (FDR < 0.01, hypergeometric test; Fig. 3a–d, Extended Data Fig. 7c, d). These observations were robust to a randomization analysis in which three female and three male subjects were chosen at random from each pathological category for 100 trials (FDR < 0.01, Extended Data Fig. 7d, Supplementary Table 5). Consistent differences were also reflected in the enrichment of quantitative pathological features (Fig. 3f, FDR < 0.01, permutation test). Thus, both categorical assignments and unbiased direct measurements of clinico-pathological variables support a strong association between the cell-type subpopulations and AD pathological status that is robust to individual selection.

To gain further insight into the molecular processes that distinguish these subpopulations, we identified marker genes for each cellular subpopulation (Extended Data Fig. 8 and Supplementary Table 6, FDR < 0.01, two-sided Wilcoxon rank-sum test).

AD-pathology-associated Ex4 neurons were marked by *LINGO1*, *RASGEF1B*, and *SLC26A3*, suggesting that subpopulations preferentially observed in neuropathology may underlie the differences in gene expression that were observed at the cell-type level. AD-pathology-associated Oli0 cells were marked by *CADM2*, *QDPR*, *NLGN1*, and *CRYAB*; the latter is an anti-apoptotic and neuroprotective chaperone, the dysfunction of which could exacerbate inflammation and demyelination⁴¹. Immunohistochemistry confirmed that cell subsets of the oligodendrocyte lineage (oligodendrocytes and oligodendrocyte precursor cells) express high levels of *CRYAB* or *QDPR* in the white matter of AD-pathology individuals (Fig. 3g, h, Extended Data Fig. 9a, b). The AD-pathology-associated astrocyte subpopulation Ast1 showed preferential expression of *GLUL* and of the AD risk factor *CLU*, which has been shown to be upregulated in reactive astrocytes in response to neurodegeneration⁴². The AD-pathology-associated subpopulations In0, Opc1, and Mic1 were marked by genes that have roles in protein folding and stability, neuronal and necrotic death, and T cell activation and immunity, respectively, suggesting cell-type-specific responses to global cellular stress (Extended Data Fig. 8). Thus, in addition to the previously reported roles of neurons and microglia in AD pathophysiology, the disease-associated signatures of oligodendrocytes and astrocytes reveal additional glial transcriptional responses to pathology.

Previous studies have profiled microglia from mouse models of AD^{23,25}. We tested whether the expression signatures associated with AD pathology that we found in the human brain overlapped with the reported states in mice. We found that marker genes of the AD-pathology-associated Mic1 subpopulation, including the MHC-II genes *CD74* and *HLA-DRB1*, significantly overlapped ($P \leq 0.01$; one-sided Fisher's exact test) with mouse disease-associated²⁵ and mouse late-response microglia²³ marker genes (Extended Data Fig. 9d, Supplementary Table 7). Immunohistochemistry of samples from AD-pathology individuals confirmed the presence of a subpopulation of microglia that expressed high levels of MHC class II proteins (Extended Data Fig. 9c). The microglial subpopulation identified here in humans revealed AD-associated genes not seen in the animal models, including the complement component *C1QB* and the pattern recognition receptor *CD14*. Next, we tested to what extent Mic1 markers also overlapped with human non-AD, aged microglia⁴³. Although we found a significant overlap ($P \leq 0.01$; Fisher's exact test, Extended Data Fig. 9d), many Mic1 marker genes, including *APOE*, were specific to AD pathology and not identified in aged microglia. Our observations suggest that the Mic1 subpopulation represents a distinct microglial state that shares features with, but is also distinct from, previously reported microglial cell states in mouse models. However, more extensive single-cell references of glial cells in the human brain will be required to contextualize pathological versus normal heterogeneity.

Sex-specific differential response to AD pathology

We identified robust differences in the association of AD pathology between cells from female versus male individuals (Fig. 3d)—AD-pathology-associated cell subpopulations (Ex4, Ast1, Oli0, and Mic1) were enriched with female cells, whereas no-pathology subpopulations (Ex6, Ast0, and Oli1) were enriched with male cells. This overrepresentation of female cells was not a result of the disproportionate cell contribution of particular individuals (Extended Data Fig. 7a, b), nor was it owing to more severe AD pathology in female individuals (Extended Data Fig. 10a). Differences between sexes were also reflected in the marker genes of AD-pathology-associated subpopulations; these genes showed expression patterns that partially segregated female and male AD individuals, with higher expression in females (Extended Data Fig. 10b, c). We thus hypothesized that the differences might stem from a sex-specific differential transcriptional response to AD pathology.

To discern whether female and male individuals present global differential responses to AD pathology, and whether such variability preferentially involves specific cell types, we recomputed individual-level gene–trait correlations, splitting the dataset by sex (Methods). The resulting data enable correlation distributions to be directly compared

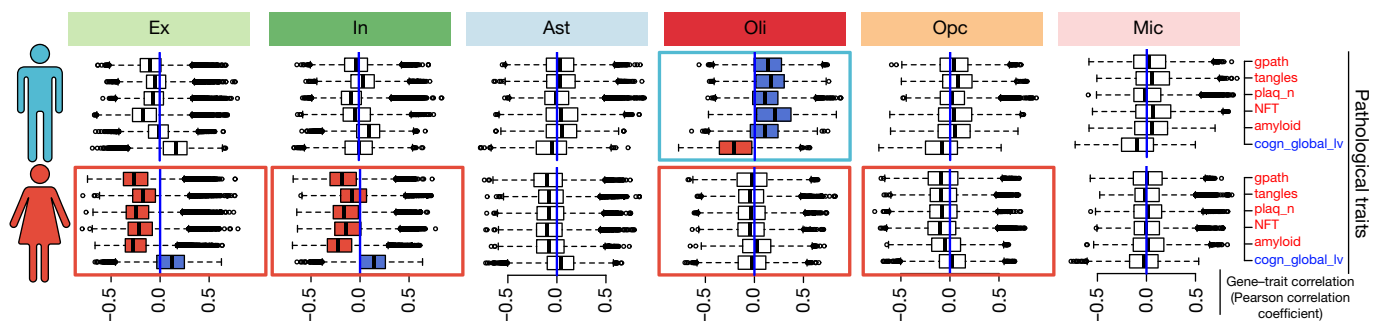


Fig. 4 | Sex-specific differential response to AD pathology. Individual-level transcriptome-wide gene–trait correlation analysis. Box plots show the distribution of correlation values (Pearson correlation coefficient) computed between gene-expression profiles ($n = 17,926$ genes) averaged for cells of each type across each individual, and the corresponding pathological measurements across individuals ($n = 24$ female individuals and $n = 24$ male individuals). Box plots are centred around the median, with the interquartile range (IQR) defining the box. The upper whisker

extends to the largest value no further than $1.5 \times \text{IQR}$ from the end of the box. The lower whisker extends to the smallest value at most $1.5 \times \text{IQR}$ from the end of the box. Pathological traits are represented in red font and cognition in blue. Rectangles around plots highlight the most contrasting differences observed between male and female transcriptional responses; blue indicates a dominant positive correlation and red a dominant negative correlation.

by cell type and pathological feature (Fig. 4). Notably, we found contrasting, qualitatively distinct responses between sexes, involving multiple cell types (Supplementary Table 8). The most extreme differences were observed in neurons and oligodendrocytes. In males, increased pathology correlated with a global transcriptional activation in oligodendrocytes, as reflected in positive correlation shifts (median correlations, 0.203 (neurofibrillary tangle burden (NFT)); 0.165 (tangles); 0.108 (amyloid); and 0.101 (plaques)) (Fig. 4, top). By contrast, global shifts in transcription in oligodendrocytes were not observed in females; instead, correlations with pathological traits and with cognition remained centred around zero (median correlations, -0.04 (NFT); -0.042 (tangles); 0.028 (amyloid); and -0.031 (plaques)) (Fig. 4). Extreme responses were also observed in neuronal cells, in particular in response to amyloid. In females, increased pathology correlated with a global downregulation (median negative correlation) of gene activity in both excitatory neurons (median correlation, -0.272 (amyloid)) and inhibitory neurons (median correlation, -0.225 (amyloid)). In males, excitatory neurons showed a qualitatively similar but much less pronounced response (median correlation, -0.01 , amyloid). Conversely, inhibitory neurons did not, in males, show a clear overall shift in response to pathology or a correlation with cognition, except for a slight increase in response to amyloid (median correlation, 0.1 (amyloid)). We also observed differences, albeit more subtle, in other cell types. In particular, oligodendrocyte precursor cells in females presented a shift towards downregulation in response to pathology—as reflected by a shift in the directionality of correlation for all pathological variables—which was not observed in males. Estimations of median sex differences (bootstrap-estimated) that support these observations are provided in Supplementary Table 8.

To examine a biological role of possible sex bias in relation to white-matter pathology, we looked at the total volume of white-matter hyperintensities (WMH) from MRI data from 505 individuals in the ROSMAP cohorts. We found a significant association between the volume of white-matter lesions and lower cognition (cogn_global_lv) in female subjects, but not in males (Extended Data Fig. 10d, e). These observations are consistent with a scenario of reduced transcriptional response, particularly in oligodendrocytes, in females with AD pathology. Although these analyses support an underlying sex bias in molecular processes linked to white-matter changes in AD pathophysiology, further single-cell-level analyses of larger sample size and additional experimental follow-up studies will be needed to fully understand the relationship between transcriptional and pathological differences between sexes.

Discussion

We report here 80,660 single-cell transcriptomes across 48 men and women with a range of AD-associated pathology. Whereas previous

studies on AD pathophysiology focused primarily on neurons and microglia (and bulk RNA is dominated by expression signals of neurons and oligodendrocytes), here we provide pathology-responsive transcriptional signatures across six major cell types—excitatory neurons, inhibitory neurons, astrocytes, oligodendrocytes, oligodendrocyte precursor cells, and microglia—as well as 40 transcriptionally distinct subpopulations of cells, some of which are preferentially overrepresented in AD pathology and differentially represented between sexes.

Although most genes presented distinct cell-type-specific perturbations, many of the top DEGs were involved in related processes across cell types. Myelination-related genes were recurrently perturbed not only in oligodendrocytes and oligodendrocyte precursor cells, but also in cells of most major cell types—possibly indicating a major regulatory response to maintain myelin integrity. Although white-matter pathologies have been documented in AD⁴⁴, single-cell resolution enabled us to identify regulators of myelination—such as *LINGO1*—that are perturbed across neuronal and glial cells, as well as other myelin-related genes that are perturbed only in neurons (*PRNP*, *CNTNAP2*, *ERBIN*, *NEGR1*, and *BEX1*) or only in glial cells (*CRYAB*). Cell-type-specific regulatory complexity may thus need to be taken into consideration in the design of therapeutic interventions in AD, and genes with a more homogeneous response, such as *LINGO1*, might have more potential for intervention²⁶.

Whereas perturbations in gene expression were largely cell-type specific at an early stage of pathology, genes upregulated in late-stage pathology tended to be common across cell types and were associated with a global stress response. Previous studies at the bulk-tissue level have implicated the downregulation of aggregation-prone proteins as a mechanism to mitigate compromised protein homeostasis in AD⁴⁵. The global cell-agnostic upregulation of proteostasis pathways observed herein may similarly constitute an intrinsic adaptive response in an attempt to balance impaired protein homeostasis in late-stage pathology. Alternatively, it might reflect a disruption of the base-level functioning of the proteostasis network—a factor known to contribute to disease progression^{34–36}.

We found that transcriptional alterations seemed to stem from changes in cell state, with certain cell-type subpopulations more readily captured in AD pathology. The observed alterations are consistent with a scenario in which existing subtypes of the normal brain are preferentially responsive to pathology and overexpress a set of responsive genes in addition to constitutive markers of cell subtypes. However, changes in the relative abundance of subtypes as a consequence of differential vulnerability to pathology is an alternative explanation that we currently cannot rule out. We observed that cells isolated from female individuals were overrepresented in many of the AD-pathology-associated cell subpopulations; responses between sexes were contrasting and qualitatively distinct, particularly in neurons and

oligodendrocytes. Our data highlight myelination-related processes in AD pathogenesis, and provide support for a sexual dimorphism in AD that manifests at the transcriptional level, even for individuals matched for age and pathology. The transcriptional sex differences may be interpreted as a greater transcriptional disease burden in female individuals. Alternatively, our findings may be viewed as indicating greater resilience in the female individuals in our study; that is, despite strong alterations at the transcriptional level, the degree of pathophysiological and cognitive decline remains similar in females and males.

Overall, our single-cell-resolution analysis highlights the complexity of glial-neuronal interactions in response to AD pathology. Building on our data, future research could investigate how to distinguish between neuroprotection versus pathogenicity, and the responsive versus the driving nature of the transcriptional alterations observed. The mechanistic basis of the observed changes during the course of AD progression remains unknown.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-019-1195-2>.

Received: 23 July 2018; Accepted: 24 April 2019;

Published online 1 May 2019.

- Masters, C. L. et al. Alzheimer's disease. *Nat. Rev. Dis. Primers* **1**, 15056 (2015).
- Hardy, J. & Selkoe, D. J. The amyloid hypothesis of Alzheimer's disease: progress and problems on the road to therapeutics. *Science* **297**, 353–356 (2002).
- Braak, H. & Braak, E. Neuropathological staging of Alzheimer-related changes. *Acta Neuropathol.* **82**, 239–259 (1991).
- De Strooper, B. & Karran, E. The cellular phase of Alzheimer's disease. *Cell* **164**, 603–615 (2016).
- Canter, R. G., Penney, J. & Tsai, L.-H. The road to restoring neural circuits for the treatment of Alzheimer's disease. *Nature* **539**, 187–196 (2016).
- Heneka, M. T. et al. Neuroinflammation in Alzheimer's disease. *Lancet Neurol.* **14**, 388–405 (2015).
- Bishop, N. A., Lu, T. & Yankner, B. A. Neural mechanisms of ageing and cognitive decline. *Nature* **464**, 529–535 (2010).
- Pimenova, A. A., Raj, T. & Goate, A. M. Untangling genetic risk for Alzheimer's disease. *Biol. Psychiatry* **83**, 300–310 (2018).
- Fisher, D. W., Bennett, D. A. & Dong, H. Sexual dimorphism in predisposition to Alzheimer's disease. *Neurobiol. Aging* **70**, 308–324 (2018).
- Nativio, R. et al. Dysregulation of the epigenetic landscape of normal aging in Alzheimer's disease. *Nat. Neurosci.* **21**, 497–505 (2018).
- Bossers, K. et al. Concerted changes in transcripts in the prefrontal cortex precede neuropathology in Alzheimer's disease. *Brain* **133**, 3699–3723 (2010).
- Zhang, B. et al. Integrated systems approach identifies genetic nodes and networks in late-onset Alzheimer's disease. *Cell* **153**, 707–720 (2013).
- Miller, J. A., Woltjer, R. L., Goodenbour, J. M., Horvath, S. & Geschwind, D. H. Genes and pathways underlying regional and cell type changes in Alzheimer's disease. *Genome Med.* **5**, 48 (2013).
- Gjoneska, E. et al. Conserved epigenomic signals in mice and humans reveal immune basis of Alzheimer's disease. *Nature* **518**, 365–369 (2015).
- Habib, N. et al. Massively parallel single-nucleus RNA-seq with DroNc-seq. *Nat. Methods* **14**, 955–958 (2017).
- Lake, B. B. et al. Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of the human brain. *Science* **352**, 1586–1590 (2016).
- Zhong, S. et al. A single-cell RNA-seq survey of the developmental landscape of the human prefrontal cortex. *Nature* **555**, 524–528 (2018).
- Lake, B. B. et al. Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. *Nat. Biotechnol.* **36**, 70–80 (2018).
- Macosko, E. Z. et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).
- Bennett, D. A. et al. Religious Orders Study and Rush Memory and Aging Project. *J. Alzheimers Dis.* **64**, S161–S189 (2018).
- He, Z. et al. Comprehensive transcriptome analysis of neocortical layers in humans, chimpanzees and macaques. *Nat. Neurosci.* **20**, 886–895 (2017).
- Lin, Y.-T. et al. APOE4 causes widespread molecular and cellular alterations associated with Alzheimer's disease phenotypes in human iPSC-derived brain cell types. *Neuron* **98**, 1141–1154 (2018).
- Mathys, H. et al. Temporal tracking of microglia activation in neurodegeneration at single-cell resolution. *Cell Reports* **21**, 366–380 (2017).
- Krasemann, S. et al. The TREM2-APOE pathway drives the transcriptional phenotype of dysfunctional microglia in neurodegenerative diseases. *Immunity* **47**, 566–581 (2017).
- Keren-Shaul, H. et al. A unique microglia type associated with restricting development of Alzheimer's disease. *Cell* **169**, 1276–1290 (2017).
- Fernandez-Enright, F. & Andrews, J. L. Lingo-1: a novel target in therapy for Alzheimer's disease? *Neural Regen. Res.* **11**, 88–89 (2016).
- Mi, S. et al. LINGO-1 negatively regulates myelination by oligodendrocytes. *Nat. Neurosci.* **8**, 745–751 (2005).
- Liang, C. et al. Erbin is required for myelination in regenerated axons after injury. *J. Neurosci.* **32**, 15169–15180 (2012).
- Scott, R. et al. Loss of *Cntnap2* causes axonal excitability deficits, developmental delay in cortical myelination, and abnormal stereotyped motor behavior. *Cereb. Cortex* **29**, 586–597 (2019).
- Poplawski, G. H. D. et al. Adult rat myelin enhances axonal outgrowth from neural stem cells. *Sci. Transl. Med.* **10**, eal2563 (2018).
- Khazaei, M. R. et al. Bex1 is involved in the regeneration of axons after injury. *J. Neurochem.* **115**, 910–920 (2010).
- Seiradake, E. et al. Structural basis for cell surface patterning through NetrinG-NG2 interactions. *EMBO J.* **30**, 4479–4488 (2011).
- Holtzman, D. M., Morris, J. C. & Goate, A. M. Alzheimer's disease: the challenge of the second century. *Sci. Transl. Med.* **3**, 77sr1 (2011).
- Andrade, W. A. et al. Early endosome localization and activity of RasGEF1b, a toll-like receptor-inducible Ras guanine-nucleotide exchange factor. *Genes Immun.* **11**, 447–457 (2010).
- Balch, W. E., Morimoto, R. I., Dillin, A. & Kelly, J. W. Adapting proteostasis for disease intervention. *Science* **319**, 916–919 (2008).
- Labbadia, J. & Morimoto, R. I. The biology of proteostasis in aging and disease. *Annu. Rev. Biochem.* **84**, 435–464 (2015).
- Kohonen, T. The self-organizing map. *Proc. IEEE* **78**, 1464–1480 (1990).
- Karch, C. M. & Goate, A. M. Alzheimer's disease risk genes and mechanisms of disease pathogenesis. *Biol. Psychiatry* **77**, 43–51 (2015).
- Davies, G. et al. Study of 300,486 individuals identifies 148 independent genetic loci influencing general cognitive function. *Nat. Commun.* **9**, 2098 (2018).
- Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech.* **2008**, P10008 (2008).
- Ousman, S. S. et al. Protective and therapeutic role for α B-crystallin in autoimmune demyelination. *Nature* **448**, 474–479 (2007).
- Shin, Y.-J. et al. Clusterin enhances proliferation of primary astrocytes through extracellular signal-regulated kinase activation. *Neuroreport* **17**, 1871–1875 (2006).
- Olah, M. et al. A transcriptomic atlas of aged human microglia. *Nat. Commun.* **9**, 539 (2018).
- Caso, F. et al. White matter degeneration in atypical Alzheimer disease. *Radiology* **277**, 162–172 (2015).
- Ciryam, P. et al. A transcriptional signature of Alzheimer's disease is associated with a metastable subproteome at risk for aggregation. *Proc. Natl Acad. Sci. USA* **113**, 4753–4758 (2016).

Acknowledgements We thank the study participants and staff of the Rush Alzheimer's Disease Center; T. F. Andreassen for technical assistance; and S. J. Barker for discussions and comments. This work was supported in part by the Cure Alzheimer's Fund (CAF), the JBP Foundation and by NIH grants RF1AG054321, RF1AG062377, RF1AG054012, U01NS110453, R01AG062335, and R01AG058002 (L.-H.T.); P30AG10161, R01AG15819, R01AG17917, U01AG46152, and R01AG57473 (D.A.B.); and U01NS110453, R01AG062335, R01AG058002, R01MH109978, R01HG008155, RF1AG054012, RF1AG062377, and U01MH119509 (M.K.). H.M. was supported by an Early Postdoc Mobility fellowship from the Swiss National Science Foundation (P2BSP3_151885).

Reviewer information Nature thanks Hongjun Song and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions This study was designed by H.M., J.D.-V., D.A.B., M.K., and L.-H.T., and directed and coordinated by M.K. and L.-H.T. H.M., Z.P., M.M., F.A., X.J., and A.J.M. performed the experiments. M.M. performed the RNAscope experiment under the supervision of B.P.H. H.M. and J.D.-V. performed the bioinformatics analysis with help from F.G., S.M., and L.H. H.M., J.D.-V., J.Z.Y., R.M.R., D.A.B., M.K., and L.-H.T. wrote the manuscript.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-019-1195-2>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-019-1195-2>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to M.K. or L.-H.T.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

METHODS

Data reporting. No statistical methods were used to predetermine sample size.

Selection of individuals from ROSMAP. We selected a total of 48 individuals from the ROS, a longitudinal cohort study of ageing and dementia in elderly nuns, priests, and brothers. The study includes clinical data collected annually, detailed post-mortem pathological evaluations, and extensive genetic, epigenomic, transcriptomic, proteomic, and metabolomic bulk-tissue profiling²⁰. For analyses in this paper, we took into account a total of 12 clinical, cognitive, and pathological hallmarks of AD identified through the ROS (Supplementary Table 3; ROS clinico-pathological variables can be found in the Supplementary Information). We selected 24 control individuals with no or very little pathology (no-pathology) and 24 age-matched individuals with a spectrum of mild to severe β -amyloid and other pathologies (AD-pathology) (Supplementary Table 1). The no-pathology group included some individuals with clinical cognitive impairment, as we selected study participants solely based on pathology and let all other variables freely associate. Details of the clinical and pathological data collection methods have been previously reported^{46–50}. Individuals were balanced between sexes (12 each), and matched for age (medians 86.7 (AD-pathology) and 87.1 (no-pathology)) and years of education (medians 19.5 (AD-pathology) and 18 (no-pathology)). Informed consent was obtained from each subject, and the Religious Orders Study and Rush Memory and Aging Project were approved by an Institutional Review Board (IRB) of Rush University Medical Center.

Isolation of nuclei from frozen post-mortem brain tissue. The protocol for the isolation of nuclei from frozen post-mortem brain tissue was adapted from a previous study⁵¹. All procedures were carried out on ice or at 4°C. In brief, post-mortem brain tissue was homogenized in 2 ml homogenization buffer (320 mM sucrose, 5 mM CaCl_2 , 3 mM $\text{Mg}(\text{CH}_3\text{COO})_2$, 10 mM Tris HCl pH 7.8, 0.1 mM EDTA pH 8.0, 0.1% IGEPAL CA-630, 1 mM β -mercaptoethanol, and 0.4 U μl^{-1} recombinant RNase inhibitor (Clontech)) using a Wheaton Dounce tissue grinder (10 strokes with the loose pestle). Homogenization buffer (3 ml) was added (to a final volume of 5 ml) and the homogenized tissue was incubated on ice for 5 min. Then the homogenized tissue was filtered through a 40- μm cell strainer, mixed with an equal volume of working solution (50% OptiPrep density gradient medium (Sigma-Aldrich), 5 mM CaCl_2 , 3 mM $\text{Mg}(\text{CH}_3\text{COO})_2$, 10 mM Tris HCl pH 7.8, 0.1 mM EDTA pH 8.0, 1 mM β -mercaptoethanol) and loaded on top of an OptiPrep density gradient (10 ml 29% OptiPrep solution (29% OptiPrep density gradient medium, 134 mM sucrose, 5 mM CaCl_2 , 3 mM $\text{Mg}(\text{CH}_3\text{COO})_2$, 10 mM Tris HCl pH 7.8, 0.1 mM EDTA pH 8.0, 1 mM β -mercaptoethanol, 0.04% IGEPAL CA-630, and 0.17 U μl^{-1} recombinant RNase inhibitor) on top of 5 ml 35% OptiPrep solution (35% OptiPrep density gradient medium, 96 mM sucrose, 5 mM CaCl_2 , 3 mM $\text{Mg}(\text{CH}_3\text{COO})_2$, 10 mM Tris HCl pH 7.8, 0.1 mM EDTA pH 8.0, 1 mM β -mercaptoethanol, 0.03% IGEPAL CA-630, and 0.12 U μl^{-1} recombinant RNase inhibitor)). The nuclei were separated by ultracentrifugation using an SW32 rotor (20 min, 9,000 r.p.m., 4°C). A total of 3 ml of nuclei was collected from the 29%/35% interphase and washed with 30 ml of PBS containing 0.04% BSA. The nuclei were centrifuged at 300g for 3 min (4°C) and washed with 20 ml of PBS containing 0.04% BSA. Then the nuclei were centrifuged at 300g for 3 min (4°C) and resuspended in 500 μl PBS containing 0.04% BSA. The nuclei were counted and diluted to a concentration of 1,000 nuclei per microlitre in PBS containing 0.04% BSA.

Droplet-based snRNA-seq. For droplet-based snRNA-seq, libraries were prepared using the Chromium Single Cell 3' Reagent Kits v2 according to the manufacturer's protocol (10x Genomics). The generated scRNA-seq libraries were sequenced using NextSeq 500/550 High Output v2 kits (150 cycles).

Analysis of droplet-based snRNA-seq data. Gene counts were obtained by aligning reads to the hg38 genome (GRCh38.p5 (NCBI:GCA_000001405.20) using CellRanger software (v2.0.0) (10x Genomics). To account for unspliced nuclear transcripts, reads mapping to pre-mRNA were counted. After quantification of pre-mRNA using the CellRanger count pipeline on each of the 48 individual libraries, the CellRanger aggr pipeline was used to aggregate all libraries and equalize the read depth between libraries before data merging (with the default parameters) to generate a gene-count matrix. Then, a cut-off value of 200 unique molecular identifiers (UMIs) was used to select single cells for further analysis. From our pilot sample analysis, we realized that the default cell-detection method used by 10x Genomics (assuming UMI values—a reflection of the RNA content—vary by roughly an order of magnitude among cells) failed to detect a large fraction of the microglia population. Therefore, to determine a more appropriate UMI cut-off value, we plotted a histogram showing cell density as a function of UMI values. On the basis of this analysis we determined 200 UMIs as the lower cut-off for cell filtering. This resulted in an initial dataset that was then further examined to exclude low-quality libraries (see 'Quality control for cell inclusion', below).

External data sources. Markers of human cortical layers and marker genes of cell-type clusters were obtained from ref. ²¹ and ref. ¹⁸, respectively, and microglia cell-state signatures were obtained from previously published studies^{23,25}.

Quality control for cell inclusion. The initial dataset contained 80,660 cells, with a median value of 1,496 total read counts over protein-coding genes. As initial reference, the entire dataset was projected onto the two-dimensional space using *t*-distributed stochastic neighbour embedding (*t*-SNE) on the top 10 principal components. The *t*-SNE coordinates were used to visualize potential biases in apparent cell similarity caused by differential cell quality. For each cell, the following quality measures were quantified: (1) the number of genes for which at least one read was mapped (which is indicative of library complexity); (2) the total number of counts; (3) the percentage of counts mapping to the top 50 genes; and (4) the percentage of reads mapped to mitochondrial genes (which may be used to approximate the relative amount of endogenous RNA, and is commonly used as a measure of cell quality). Cells with a high ratio of mitochondrial relative to endogenous RNAs had low starting amounts of RNA, which might indicate that source cells were dead or stressed and thus result in RNA degradation. Outlier cells in these quality metrics were found to cluster together in the *t*-SNE two-dimensional space. On the basis of these observations and subsequent scatter-plot analyses, cells with fewer than 200 detected genes and cells with an abnormally high ratio of counts mapping to mitochondrial genes (relative to the total number of detected genes) were removed. Specifically, given a highly skewed empirical distribution of the mitochondrial ratio values (that is, having an elbow shape clearly separating high and low scores), outlier cells were classified in two groups using the *k*-means clustering algorithm ($k = 2$) on the mitochondrial ratio, and subsequently removed. Only counts associated with protein-coding genes were considered; mitochondrially encoded genes and genes detected in fewer than 2 cells were excluded. After applying these filtering steps, the dataset included 17,926 genes profiled in 75,060 nuclei.

Cell clustering. All 75,060 cells were combined into a single dataset. Normalization and clustering were done with the SCANPY package⁵². In brief, counts for all nuclei were scaled by the total library size multiplied by 10,000, and transformed to log space. A total of 3,188 highly variable genes were identified based on dispersion and mean, the technical influence of the total number of counts was regressed out, and the values were rescaled. These preprocessing steps were performed by sequentially using the functions `normalize_per_cell`, `filter_genes_dispersion`, `log1p.regress_out`, and `scale` in SCANPY. Principal component analysis (PCA) was performed on the variable genes, and *t*-SNE was run on the top 10 principal components (PCs) using the Multicore-TSNE package (<https://github.com/DmitryUlyanov/Multicore-TSNE>). The top 50 PCs were used to build a *k*-nearest-neighbours cell-cell graph with $k = 30$ neighbours; subsequently spectral decomposition over the graph was performed with 15 components, and the Louvain graph-clustering algorithm was applied to identify cell clusters. These analyses were performed using the functions `pca`, `neighbours`, and `louvain` in SCANPY. We confirmed that a number of PCs greater than 30 captures 100% of the variance of the data. The initial pre-clustering analysis resulted in 20 pre-clusters with a median number of 2,990 cells, ranging from 413 to 15,900 cells, after excluding 2 pre-clusters of 360 and 791 cells that reflected low-quality cells (that is, cells that showed mixed cell-type markers; extreme complexity with many more genes expressed than other cells; either too many or too few reads; and, in one case, cells isolated almost exclusively from one individual). For each pre-cluster, differentially expressed genes were detected using the variance-adjusted *t*-test as implemented in the function `rank_genes_groups` in SCANPY. The top 500 ranking genes were extracted for each cluster, and used to test for overlap with markers as previously reported¹⁸. The same clustering protocol was used for both pre- and sub-clustering analyses. During sub-clustering, additional potentially spurious clusters representing low-quality or doublet cells were detected on the basis of extreme separation from the rest of the sub-clusters from the same cell type. Of these, those having a distinctly high number of total counts and mixed expression of markers from different cell types were tagged as potential doublets and not considered for downstream analyses, resulting in a total of 70,634 cells.

Cell type annotation and sub-clustering. For each pre-cluster, we assigned a cell-type label using statistical enrichment for sets of marker genes^{18,21}, and manual evaluation of gene expression for small sets of known marker genes. Enrichment was statistically assessed using the hypergeometric distribution (Fisher's exact test) and FDR correction over all gene sets and pre-clusters. Broad cell-type clusters were defined by grouping together all pre-clusters corresponding to the same cell type. Sub-clustering analysis was performed independently over each broad cell-type cluster.

Sub-cluster-trait associations. Cell sub-clusters were annotated for phenotypic traits and AD-pathology status by assessing, within each sub-cluster, the overrepresentation of cells isolated from individuals annotated with the corresponding trait. For categorical phenotypic variables, enrichment was evaluated using the hypergeometric distribution (Fisher's exact test) and FDR correction over all gene sets and sub-clusters. Enrichment or depletion of quantitative pathological variables was assessed individually by contrasting the average observed value across the cells of a given sub-cluster with a corresponding null distribution estimated by randomly resampling sub-cluster label assignments and computing an average

score 10,000 times. The deviation of the observed value from the random expected distribution was quantified using a z -score statistic. Statistical overrepresentation analysis for cells isolated from each individual across the sub-clusters (Fisher's exact test) verified that individual sub-clusters were not exclusively composed of cells from one or a few subjects (Extended Data Fig. 7b). Robustness of sub-cluster–trait associations was assessed using a randomization study in which cells from 3 female and 3 male subjects randomly chosen from each pathological category (AD-pathology and no-pathology) were subjected to a computing enrichment analysis for 100 trials. Aggregate P value calculations were conducted by computing a meta- p -value using the R package *metap* (Extended Data Fig. 7d).

Phenotypic analysis. Post-mortem scores on AD-related pathologic indices for amyloid, Braak staging, CERAD score, cognitive diagnosis, global cognitive function, global AD-pathology burden, neurofibrillary tangle burden, NIA-Reagan score, diffuse and neuritic plaque burden, and neuronal neurofibrillary tangle density were used to quantitatively describe the pathological phenotype of each individual (Supplementary Information). The quantitative values tangles, NFT, gpath (aggregate of neuritic plaque, diffuse plaque, and neurofibrillary tangle scores), plaq_n, amyloid, and cogn_global_lv were selected as the primary markers for quantifying AD progression. To identify AD pathological groups, distance-based k -means clustering was applied to the quantitative pathological features using $k = 3$. The number of clusters was selected by testing multiple increasing values of k , and taking the value at which the within-cluster sum-of-square distance dropped.

Marker identification. For sub-clusters, a set of markers (specifically over-expressed) genes was defined by a differential expression analysis of the cells grouped in each sub-cluster against the remaining cells within the corresponding broad cell-type cluster. This analysis was applied to all cell types independently. Significantly overexpressed genes were defined based on the Wilcoxon rank-sum test with a FDR-corrected P value ≤ 0.01 and a \log_2 (mean gene expression across cells in sub-cluster/mean gene expression across cells in other sub-clusters) of 0.5. Only genes detected in at least 25% of the cells within the given sub-cluster were considered.

Gene Ontology enrichment analyses were performed using the R package *gProfileR* (<https://cran.r-project.org/web/packages/gProfileR/index.html>) and using the P value-ranked gene lists as input. For all cases we used the set of 17,926 protein-coding genes included in the quality control data as background.

Differential gene-expression analysis. snRNAseq-based differential expression analysis was assessed using two tests. First, a cell-level analysis was performed using the Wilcoxon rank-sum test and FDR multiple-testing correction. Second, a Poisson mixed model accounting for the individual of origin for nuclei and for unwanted sources of variability was performed using the R packages *lme4* and *RUV-seq*, respectively. The consistency of DEGs detected using the cell-level analysis model with those obtained with the Poisson mixed model was assessed by comparing the directionality and rank of DEGs in the two models. Consistency in directionality for all cell types was measured by counting the fraction of the top 1,000 DEGs (ranked by FDR scores) detected in cell-level analysis that showed consistent direction in the mixed model. High consistency was found, with a median fraction of 0.99 (Supplementary Table 2). Global consistency between the two models was assessed statistically using a resampling test. We tested whether the differential P value and z -score ranks corresponding to genes detected as upregulated or downregulated in the cell-level analysis were significantly higher or lower than those that were expected by chance when computed using the mixed model. Expected scores were estimated by randomly sampling same-sized gene sets ($n = 1,000$ replicates). Both significance rank and direction deviated significantly from expectation, with directionality consistent for up- or downregulated genes. Results from the consistency tests and from both the cell-level and mixed-model differential tests are included in Supplementary Table 2. For analyses involving DEG counts, only genes that were significantly supported by both models using the criteria FDR-corrected $P < 0.01$ in a two sided Wilcoxon-rank sum test, absolute \log_2 (mean gene expression in AD category x /mean gene expression in AD category y) > 0.25 , and FDR-corrected $P < 0.05$ in a Poisson mixed model were considered. Owing to their low cell counts, pericyte and endothelial cell populations were excluded from differential analyses.

Bulk RNA-seq differential analysis was performed by fitting a linear model using the R package *limma*⁵³, accounting for the covariates age, RNA integrity number, post-mortem interval, and plate batches. The pathological definition of AD groups provided by ROSMAP was used. This classification defines AD or non-AD groups based exclusively on the overall pathological burden, without considering clinical diagnosis. We used the labels AD-pathology and no-pathology for consistency.

Consistency of gene-expression perturbations in the different cell types observed in snRNA-seq data with those detected in tissue-level bulk RNA-seq was assessed using a resampling approach. To test whether the genes identified as DEGs in single-cell data were also detected as high ranking in the differential analysis in bulk data, a z -score statistic was computed to quantify the deviation of the observed differential (P value) rank scores obtained in the bulk analysis for the

genes detected as DEGs in single cells, relative to those observed in 1,000 randomly chosen gene sets (Supplementary Table 2). This analysis was performed for each cell type independently.

Correlation analysis of gene expression and AD-related neuropathological traits using SOMs. First, for each major cell type, a gene-wise correlation coefficient (Spearman's rank correlation coefficient) was computed using gene expression and AD-related neuropathological trait values across all the annotated cells as variables. The AD-related neuropathological traits included in this analysis were cogn_global_lv, age_death (age at death), educ (years of education), msex (self-reported sex), parksc_lv (global Parkinsonian summary score (last valid score)), gpath, gpath_3neocort (global measure of neocortical pathology), pmi (post-mortem interval), amyloid, plaq_d (diffuse plaque burden), plaq_n, NFT, and tangles. Only significant correlations after Bonferroni correction at $P < 0.01$ were considered. The resulting correlation matrices for each major cell type were concatenated and analysed using a computational algorithm (SOM)³⁷. All SOMs were created using the kohonen R package⁵⁴. To identify the territories of the SOM most strongly correlated with AD-related neuropathological traits, we used an image segmentation method and further manual curation to identify territories (gene–trait correlation modules) based on all the individual cell-type-specific SOM plots for each neuropathological trait. Enrichment analysis for Gene Ontology terms among the genes of a gene–trait correlation module was performed using Metascape⁵⁵. The robustness of gene–trait associations to single-cell heterogeneity and noise was confirmed examining individual-level correlations for the genes in a gene–trait module. Individual-level correlations were computed by first averaging for each individual the normalized gene-expression profiles across cells of the same cell. This resulted in cell-type-specific averaged gene-expression profiles across the 48 individuals. Average profiles were subsequently mean-centred and scaled to finally compute gene-wise correlation coefficients versus corresponding pathological values. Individual-level gene–trait correlations were computed independently for all 48 individuals, for only the 24 male individuals, and for only the 24 female individuals. The robustness of gene–trait associations to potential confounding variables was corroborated by confirming the cell-type-specific recovery of identified gene–trait modules using partial correlation. In brief, the partial Pearson's correlation coefficient between average gene expression and each pathological trait, after correcting for the effects of post-mortem interval, age, gender, and education level of each individual; was computed by first orthogonalizing the normalized expression with respect to the normalized covariates and then computing the correlation in the residual subspace.

Immunohistochemistry. Fixed human brain tissue (prefrontal cortex, BA10) was sectioned at 50 μ m using a vibratome (Leica). To retrieve the antigens, sections were incubated at 95 °C in immunohistochemistry (IHC) antigen retrieval solution (ThermoFisher Scientific; 00-4955-58) containing 0.05% Tween-20 for 45 min and then placed in PBS for 20 min at room temperature. After washing with PBS (3 \times 10 min), the brain sections were incubated in quenching solution (50 mM ammonium acetate, 100 mM CuSO₄) at room temperature overnight. After washing with double-distilled water (1 \times 15 min) and PBS (3 \times 15 min), the sections were permeabilized in PBS containing 0.3% Triton X-100 for 10 min and blocked in PBS containing 0.3% Triton X-100 and 5% normal donkey serum at room temperature for 1 h. The sections were incubated overnight (anti- β -amyloid (D54D2) antibody) or for 48 h (anti-IBA1 and anti-human HLA-DP, HLA-DQ, and HLA-DR antibodies) at 4 °C in primary antibody in PBS with 0.3% Triton X-100 and 5% normal donkey serum. Primary antibodies were anti-IBA1 (1:500; Synaptic Systems; 234 004, polyclonal guinea pig anti-serum), anti-human HLA-DP, HLA-DQ, and HLA-DR antigens, clone CR3/43 (1:100, Agilent; M077501-2), and anti- β -amyloid (D54D2) (1:500, Cell Signaling Technology; 8243). The sections were washed with PBS containing 0.1% Triton X-100 at room temperature (4 \times 15 min), and then incubated with secondary antibodies (dilution 1:2,000) overnight at 4 °C. Primary antibodies were visualized with Alexa-Fluor 488, Alexa-Fluor 594, and Alexa-Fluor 647 antibodies (Molecular Probes), and cell nuclei were visualized with Hoechst 33342 (Sigma-Aldrich; 94403). The sections were washed with PBS containing 0.1% Triton X-100 at room temperature (4 \times 15 min) and then mounted on Fisherbrand Superfrost Plus microscope slides in ProLong Gold Antifade Mountant. Images were acquired using a confocal microscope (LSM 880; Zeiss) with a 20 \times or 63 \times objective.

For tyramide signal amplification (TSA) labelling, sections were incubated at 95 °C in IHC antigen retrieval solution (ThermoFisher Scientific; 00-4955-58) containing 0.05% Tween-20 for 45 min and then placed in PBS for 20 min at room temperature. After washing with PBS (3 \times 10 min), the brain sections were incubated in quenching solution (50 mM ammonium acetate, 100 mM CuSO₄) at room temperature overnight. After washing with double-distilled water (1 \times 15 min) and PBS (3 \times 15 min), endogenous peroxidases were quenched with 0.3% H₂O₂ for 20 min at room temperature followed by washing with PBS (3 \times 10 min). Then the sections were blocked in TNB blocking buffer (0.1 M Tris HCl pH 7.5, 0.15 M NaCl, and 0.5% TSA blocking reagent (PerkinElmer; FP1020)) for 30 min at room

temperature. Primary antibody incubation was performed for 48 h at 4°C in TNB blocking buffer. The sections were washed with PBS containing 0.3% Triton X-100 at room temperature (4 × 15 min) and then incubated with horseradish peroxidase (HRP)-labelled secondary antibody (1:1,500) in TNB blocking buffer for 30 min at room temperature. Then the sections were washed with PBS containing 0.3% Triton X-100 at room temperature (4 × 15 min) and the fluorophore reaction was performed using the TSA Plus Cyanine 5 and Fluorescein System (PerkinElmer; NEL754001KT) for 3 min at room temperature. Then the sections were washed with PBS containing 0.3% Triton X-100 at room temperature (4 × 15 min) and incubated at 95°C in IHC antigen retrieval solution (ThermoFisher Scientific; 00-4955-58) containing 0.05% Tween-20 for 45 min. After washing with PBS containing 0.3% Triton X-100 at room temperature (4 × 15 min) the sections were either mounted or blocked in TNB blocking buffer for another round of TSA labelling. The following primary antibodies were used: anti-OLIG2 (1:1,000; Atlas; HPA003254; rabbit polyclonal), anti-CRYAB (1:200; LSBio; LS-B3696; rabbit polyclonal), and anti-QDPR (1:2,500; Atlas; HPA065649; rabbit polyclonal). The following secondary antibody was used: HRP-labelled anti-rabbit IgG (goat) (PerkinElmer; NEF812001EA). Images were acquired using a confocal microscope (LSM 880; Zeiss) with a 20× or 63× objective.

RNA in situ hybridization. Frozen brain tissue was embedded in Tissue-Tek OCT compound (VWR; 25608-930), cryo-sectioned to 16-μm thickness, and placed onto Fisherbrand Superfrost Plus microscope slides (ThermoFisher Scientific; 12-550-15). The RNAscope 2.5 HD Duplex Detection Kit (Chromogenic) was used according to the manufacturer's instructions, with minor modifications to tissue preparation. In brief, sections frozen at -80°C were washed with 1 × PBS, baked at 60°C for 15 min, and fixed in 10% neutral buffered formalin for 90 min at room temperature. Following fixation, the tissue was dehydrated using 50%, 70% and 100% ethanol and then baked at 60°C for 30 min to avoid detachment. Tissues were then treated with H₂O₂ for 10 min, followed by protease IV for 30 min. Brain sections were hybridized with two mRNA probes: *SLC17A7* (RNAscope probe Hs-SLC17A7-C2; ACD; 415611-C2) and *NTNG1* (RNAscope probe Hs-NTNG1; ACD; 446101). Probes in channel 1 (*NTNG1*) were labelled with HRP enzyme and visualized with a green substrate; and probes in channel 2 (*SLC17A7*) were labelled with alkaline phosphatase enzyme and a red substrate. Haematoxylin was used to mark cell nuclei in blue. Slides were then imaged using a Leica Dmi8 microscope to take 40× bright-field images. For the quantification, a blinded researcher manually counted the cells labelled with red substrate (*SLC17A7*) and determined how many of these cells were co-labelled with green substrate (*NTNG1*). The quantification was based on the analysis of BA10 tissue sections from four low-amyloid individuals and four high-amyloid individuals, with five to six images per individual.

Isolation of neuronal nuclei, cDNA synthesis and RT-qPCR analysis. Post-mortem brain tissue was homogenized in 2 ml NF1 buffer (0.5% Triton X-100, 0.1 M sucrose, 5 mM MgCl₂, 1 mM EDTA, 10 mM Tris HCl pH 8.0, 1 mM β-mercaptoethanol, 0.4 U μl⁻¹ recombinant RNase inhibitor (Clontech)) using a Wheaton Dounce tissue grinder (15 strokes with the loose pestle). After homogenization, an additional 8 ml of NF1 buffer was added and the samples were inverted 10 times to mix. The homogenate was passed through a 40-μm cell strainer (VWR; 21008-949) and centrifuged at 300g for 3 min at 4°C. Then the supernatant was removed and the nuclei were incubated with Alexa-Fluor-488-conjugated anti-NeuN antibody (Millipore Sigma; MAB377X) in PBS containing 1% BSA for 15 min at 4°C on a shaker. The nuclei were washed by adding 35 ml PBS containing 1% BSA, spun down at 300g for 3 min at 4°C, resuspended in 0.5 ml PBS containing 0.04% BSA and 0.4 U μl⁻¹ recombinant RNase inhibitor (Clontech), and stained with NucBlue Live ReadyProbes Reagent (ThermoFisher Scientific; R37605). NeuN-positive nuclei were then directly sorted into RNA lysis buffer (Qiagen, 74134) using fluorescence-activated cell sorting. Total RNA was extracted using the RNeasy Plus Mini Kit (Qiagen; 74134) according to the manufacturer's instructions, and reverse transcribed using the iScript cDNA Synthesis Kit (BioRad; 170-8891). For gene-expression analysis, cDNA was quantitatively amplified on a thermal cycler (BioRad) using SsoFast EvaGreen Supermix (BioRad; 1725202) and gene-specific primers (*DHFR* forward, GCTGGAGTATTGATCCCGCC, *DHFR* reverse, CTGGACTATGTTCCGCCAC; *STMN1* forward, ATACACT GCCTGTCGCTTGT, *STMN1* reverse, CTTTGTACCGAGGGCTGAGA; *MTRNR2L8* forward, ATTGACCTGCCCCGTGAAGAG, *MTRNR2L8* reverse, AGGGCCTGTGGACTTGTTAAG; *LINGO1* forward, ACCGCATC AAAACGCTCAAC, *LINGO1* reverse, CTAGCGGGATGAGCTTCAGG; *BEX3* forward, ATGGGCCATACCAATAGGC, *BEX3* reverse, AGAGAGCTCC CCCATAAGGA; *BEX1* forward, CGTCACTCGTGTCTCGCTAC, *BEX1* reverse, CCATTACTCCTGGGCCATCC; *RPL13* forward, CCTTCCGCTCCG CTGTTT, *RPL13* reverse, GGCCTTACGTCTGCGGATCT). The comparative C_t method was used to examine differences in gene expression and values were normalized to expression levels of *RPL13*.

Statistical analysis of white-matter data from ROSMAP cohort studies.

Measures of WMH were obtained through in vivo brain MRI data collected through the ROSMAP cohort studies. Participants undergo MRI in the Rush Alzheimer's Disease Center (RADC) cohort studies and substudies biennially on 3T scanners including the T1-weighted 3D Magnetization Prepared Rapid Acquisition Gradient Echo (MPRAGE) and T2-weighted 2D Fluid-Attenuated Inversion Recovery (FLAIR) sequences. Raw MPRAGE images were processed to generate total volumes including grey matter, white matter, cerebrospinal fluid and intracranial volumes using the statistical parametric mapping (SPM) package⁵⁶. White-matter lesions that appeared hyperintense in T2-weighted images were segmented based on FLAIR and MPRAGE data using BIANCA⁵⁷. A mask of WMH was generated and the total volume of hyperintensities calculated. Cognitive scores were taken from the recorded variable *cogn_global_lv* (Supplementary Information), which measures averaged z-scored values from a battery of 19 cognitive tests, yielding a global cognitive function summary. A negative z-score simply means that someone has an overall score that is lower than the average of the entire cohort at baseline. High- and low-cognition groups were defined based on whether a subject has an overall score lower (low-cognition, z-score <0) or higher (high-cognition, z-score >0) than the average. WMH values were compared between low- and high-cognition groups in male and female subjects, using the Wilcoxon rank-sum test (Extended Data Fig. 10d). Statistical estimation of significant difference in WMH between low-cognition and high-cognition groups in females, and between low-cognition and high-cognition groups in males, was assessed by bootstrap point and 95% confidence interval estimation of the effect size (mean difference) between groups, correcting for sample size (resampling *n* = 40 observations per group 1,000 times) (Extended Data Fig. 10e).

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

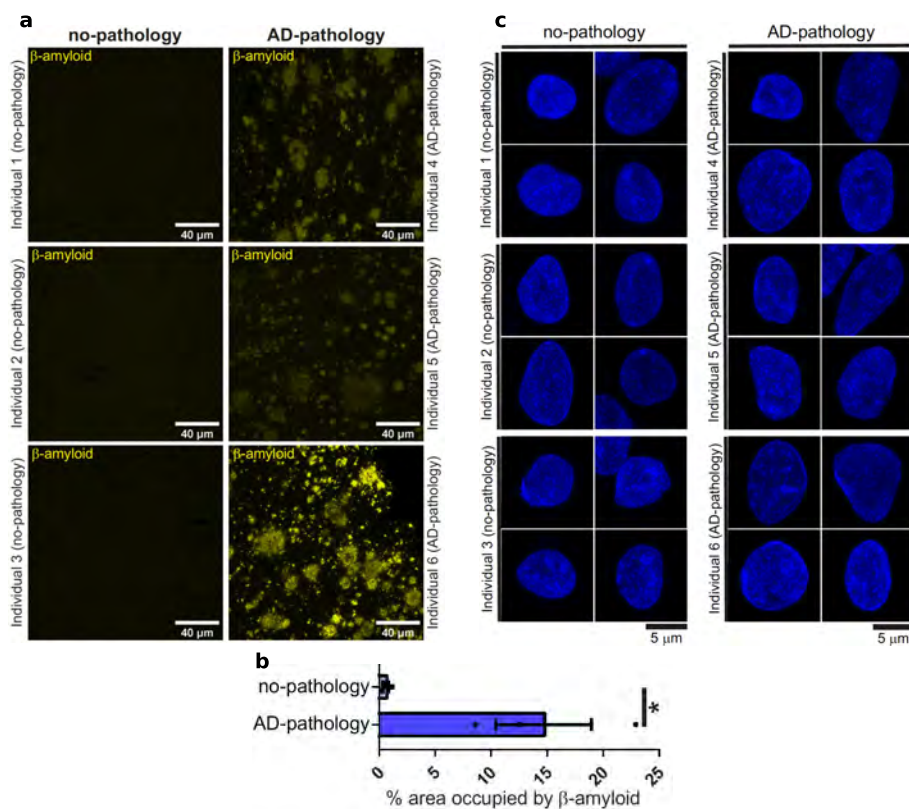
Data availability

The snRNA-seq data are available on The Rush Alzheimer's Disease Center (RADC) Research Resource Sharing Hub at <https://www.radc.rush.edu/docs/omics.htm> (snRNA-seq PFC) or at Synapse (<https://www.synapse.org/#!Synapse:syn18485175>) under the doi 10.7303/syn18485175. The ROSMAP metadata can be accessed at <https://www.synapse.org/#!Synapse:syn3157322>. The data are available under controlled use conditions set by human privacy regulations. To access the data, a data use agreement is needed. This registration is in place solely to ensure anonymity of the ROSMAP study participants. A data use agreement can be agreed with either Rush University Medical Center (RUMC) or with SAGE, who maintains Synapse, and can be downloaded from their websites.

Code availability

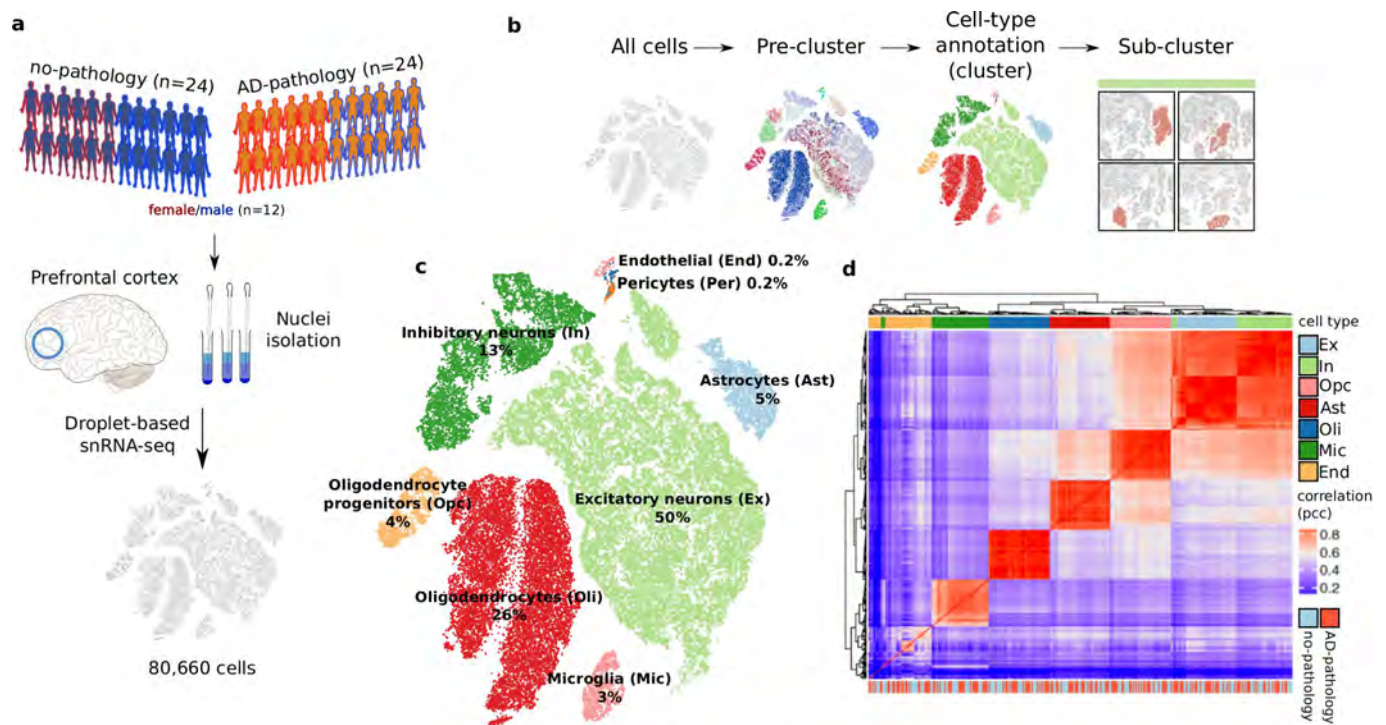
Code used throughout this study is available upon reasonable request from the corresponding authors.

- Bennett, D. A. et al. Natural history of mild cognitive impairment in older persons. *Neurology* **59**, 198–205 (2002).
- Bennett, D. A. et al. Apolipoprotein E ε4 allele, AD pathology, and the clinical expression of Alzheimer's disease. *Neurology* **60**, 246–252 (2003).
- Bennett, D. A., Schneider, J. A., Wilson, R. S., Bienias, J. L. & Arnold, S. E. Neurofibrillary tangles mediate the association of amyloid load with clinical Alzheimer disease and level of cognitive function. *Arch. Neurol.* **61**, 378–384 (2004).
- Bennett, D. A. et al. Neuropathology of older persons without cognitive impairment from two community-based studies. *Neurology* **66**, 1837–1844 (2006).
- Bennett, D. A. et al. Decision rules guiding the clinical diagnosis of Alzheimer's disease in two community-based cohort studies compared to standard practice in a clinic-based cohort study. *Neuroepidemiology* **27**, 169–176 (2006).
- Swiech, L. et al. In vivo interrogation of gene function in the mammalian brain using CRISPR-Cas9. *Nat. Biotechnol.* **33**, 102–106 (2015).
- Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
- Ritchie, M. E. et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
- Wehrens, R. & Buydens, L. M. C. Self- and super-organizing maps in R: the kohonen package. *J. Stat. Softw.* **21**, 1–19 (2007).
- Zhou, Y. et al. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat. Commun.* **10**, 1523 (2019).
- Friston, K. J. et al. Spatial registration and normalization of images. *Hum. Brain Mapp.* **3**, 165–189 (1995).
- Griffanti, L. et al. BIANCA (Brain Intensity AbNormality Classification Algorithm): a new tool for automated segmentation of white matter hyperintensities. *Neuroimage* **141**, 191–205 (2016).



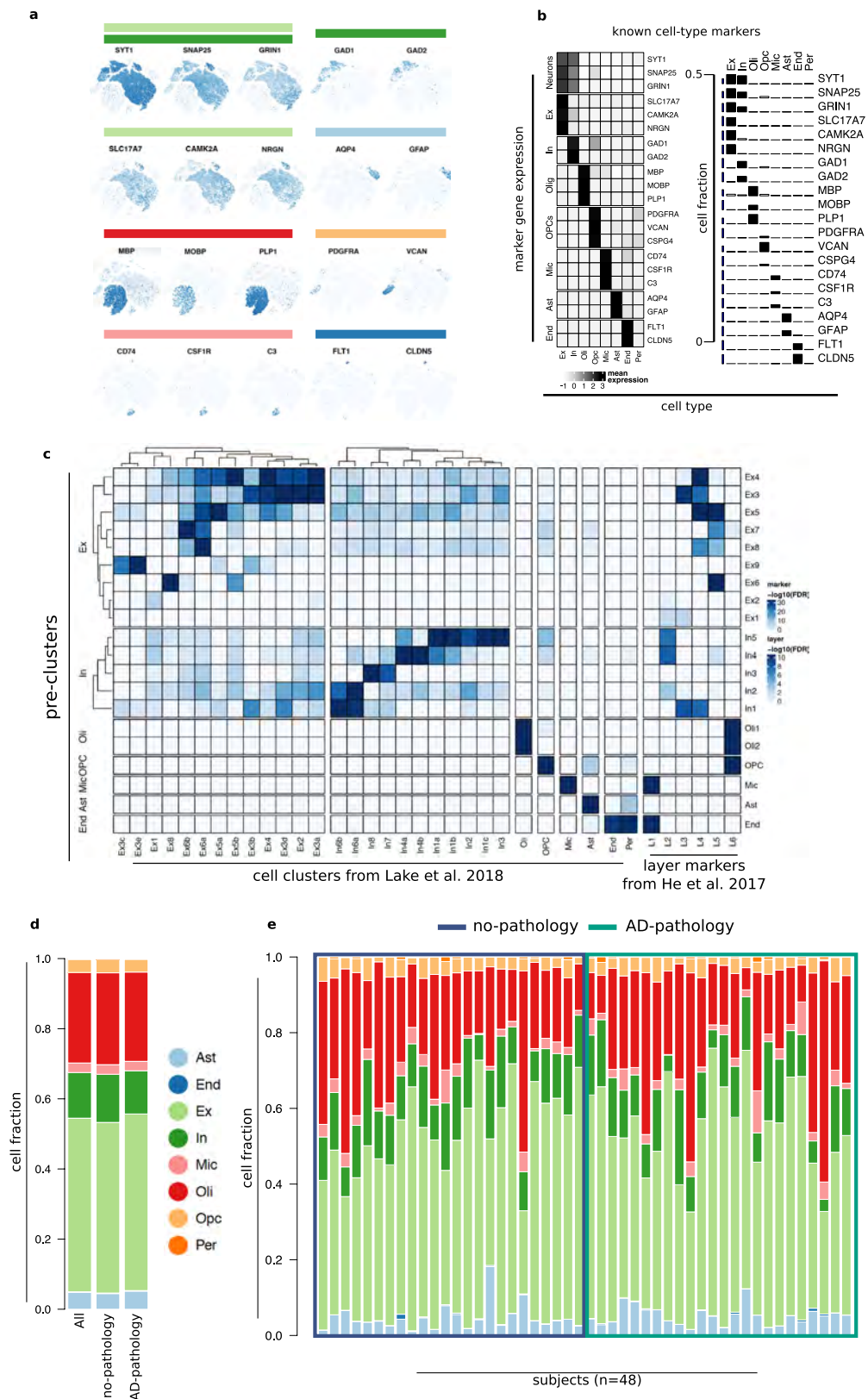
Extended Data Fig. 1 | Pathological status verification and physical integrity of isolated neuronal nuclei. a, Immunohistochemistry with anti- β -amyloid antibody (D54D2, yellow) in the grey matter of Brodmann area 10 of no-pathology and AD-pathology individuals. **b**, Quantification of the β -amyloid immunostaining in **a**. Data are mean \pm s.e.m.;

* $P = 0.030$ (Student's two-tailed t -test). **c**, High-resolution confocal microscopy images of neuronal nuclei isolated from no-pathology and AD-pathology individuals and stained with Hoechst. The experiment was performed once.



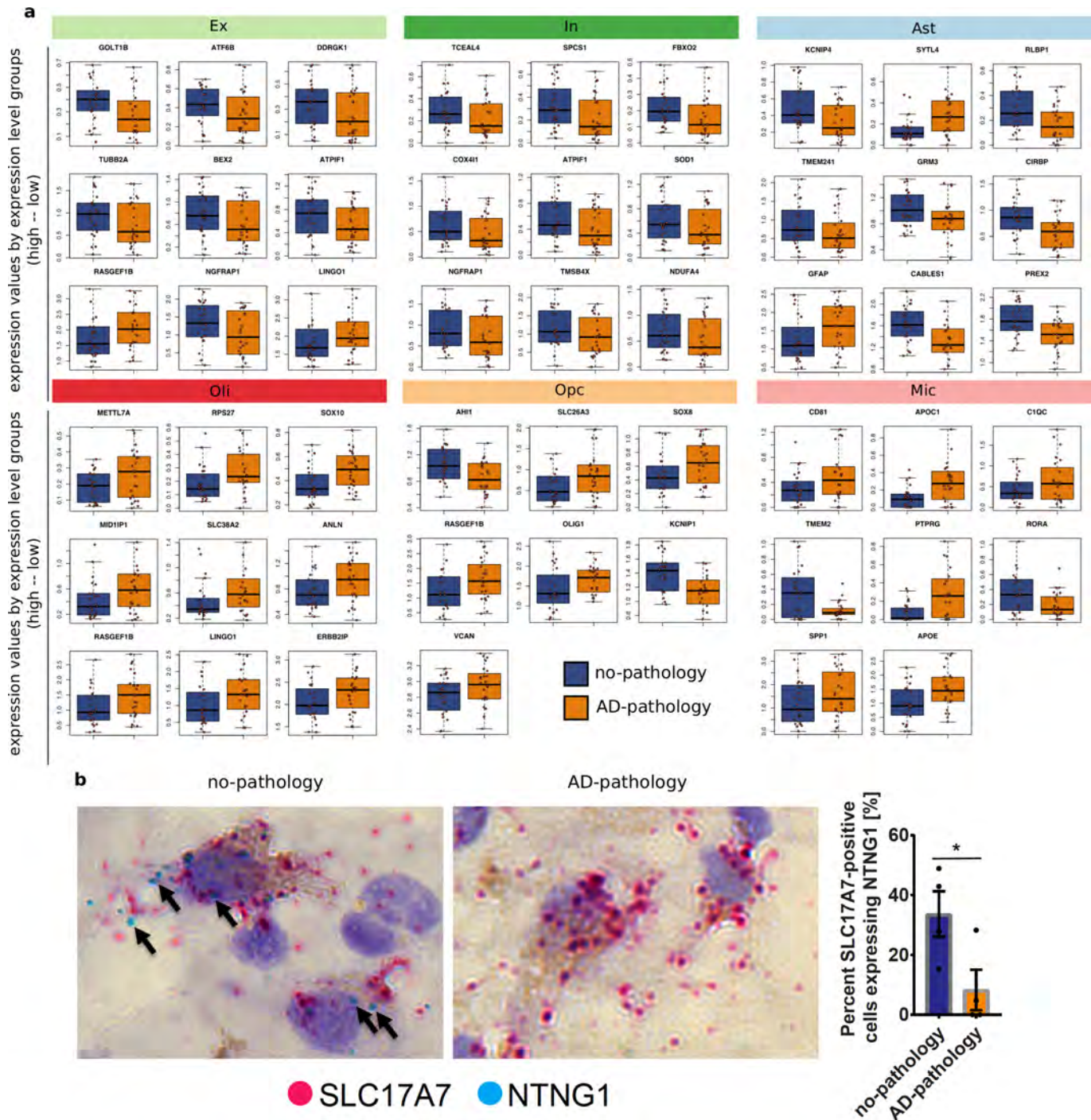
Extended Data Fig. 2 | snRNA-seq profiling and cell-type characterization. **a**, Study cohort and sample preparation. **b**, Clustering analysis workflow. **c**, Two-dimensional *t*-SNE projection of all annotated

cells ($n = 75,060$ from 24 pathology and 24 no-pathology individuals). **d**, Correlation matrix (Pearson correlation coefficient; pcc) of the average expression profiles by cell type for each individual.



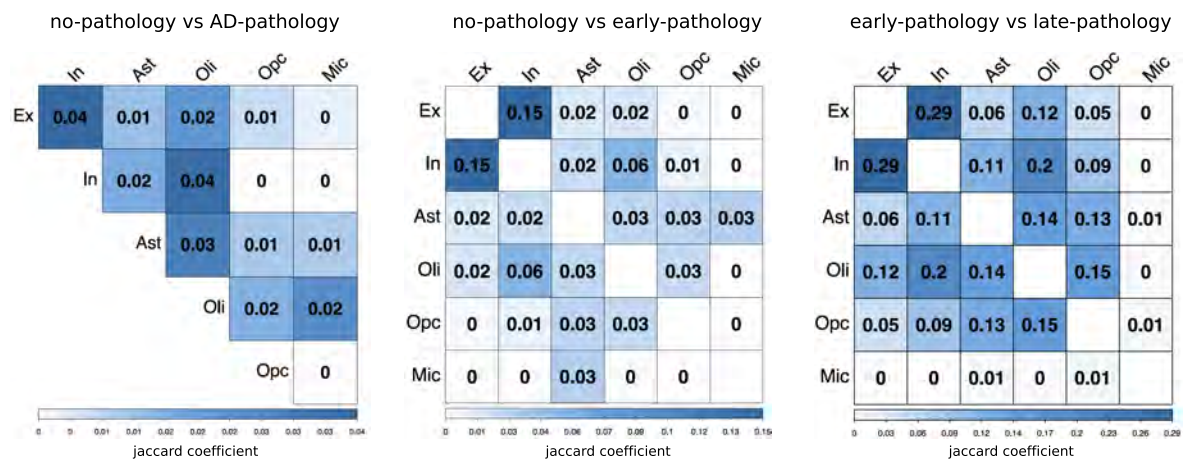
Extended Data Fig. 3 | Consistency of cells of the same type across individuals. **a**, Expression of known cell-type marker genes for each cell type. **b**, Left, expression of known cell-type marker genes in each cluster; right, fraction of cells in each cluster that express each marker gene. Vertical dashed blue line represents a scale bar referencing 0.5. **c**, Overrepresentation analysis (hypergeometric test) within each of the

pre-cluster marker sets (rows) of genes previously identified as markers¹⁸ (columns, left; $n = 1,729$ total genes) and genes previously identified as markers of cortical layers²¹ (columns, right; $n = 3,400$ total genes). **d**, Fraction of cells of each type isolated across all ($n = 48$), no-pathology ($n = 24$), and AD-pathology ($n = 24$) individuals. **e**, Fraction of cells of each type isolated from each individual (columns; $n = 48$).



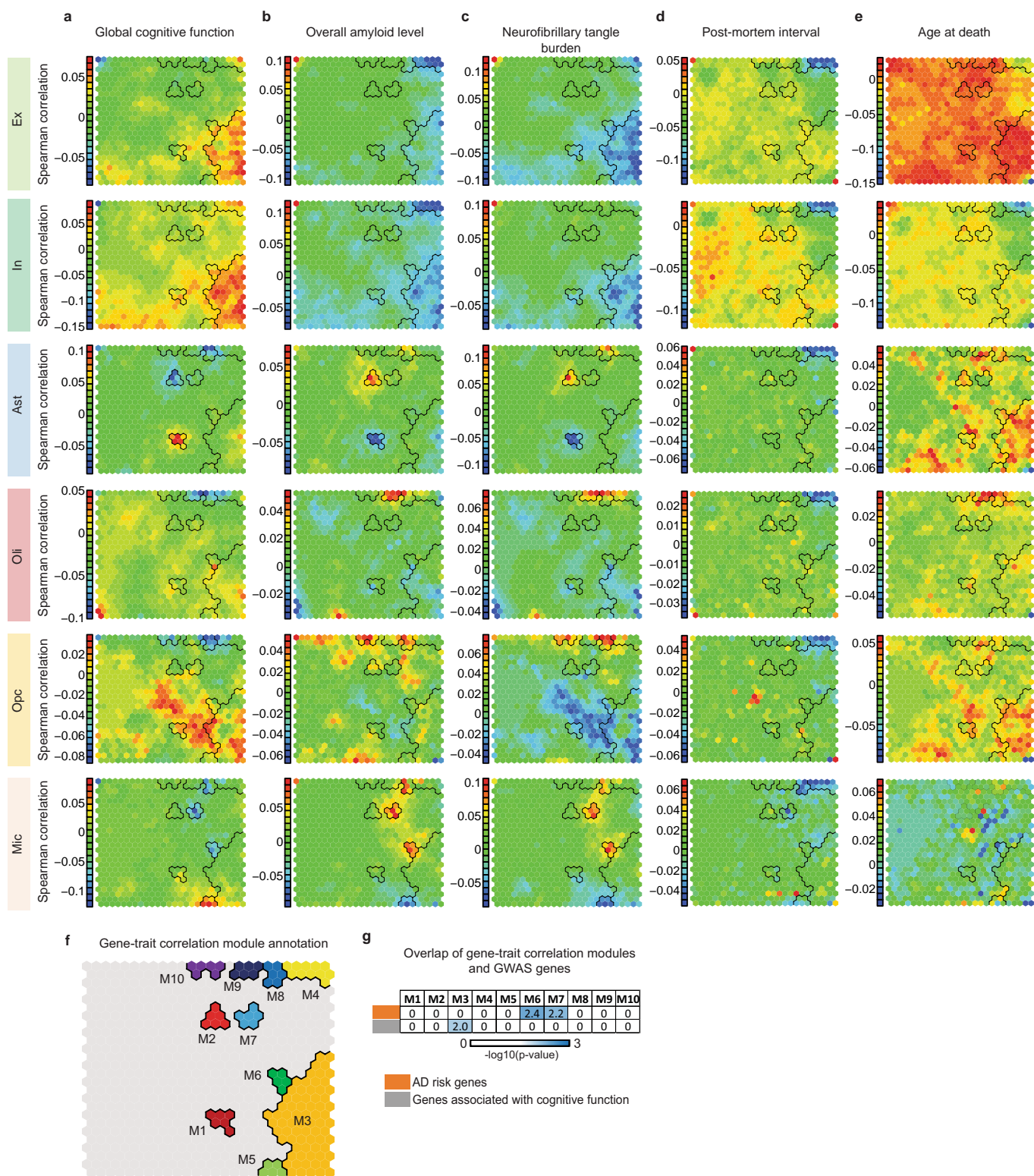
Extended Data Fig. 4 | Expression values and validation of top DEGs.
a, Mean expression values of genes across the nuclei isolated from each individual. Each point represents one individual. DEGs were classified as low, mid, or high in expression, based on their median expression level across the cells of the corresponding cell type. Groups were defined based on k -means clustering ($k = 3$). The top three genes for each group (low-, mid-, and high-expression levels) and for each cell type are shown. For oligodendrocyte precursor cells and microglia, only one and two genes, respectively, were classified within the high-expression group. *ATP1F1* is also known as *ATP5IF1*, *NGFRAP1* is also known as *BEX3*,

TMEM2 is also known as *CEMP2*, and *ERBB2IP* is also known as *ERBIN*.
b, Left, RNA in situ hybridization (RNAscope) with probes that detect the excitatory neuron marker *SLC17A7* (red) and *NTNG1* (blue) in the grey matter of Brodmann area 10 of a no-pathology and an AD-pathology individual. The tissue was counterstained with haematoxylin. Right, quantification of RNA in situ hybridization on Brodmann area 10 tissue sections. Data are mean \pm s.e.m.; * $P = 0.047$ (Student's two-tailed t -test). $n = 4$ no-pathology and $n = 4$ AD-pathology individuals; $n = 5$ or 6 images per individual.



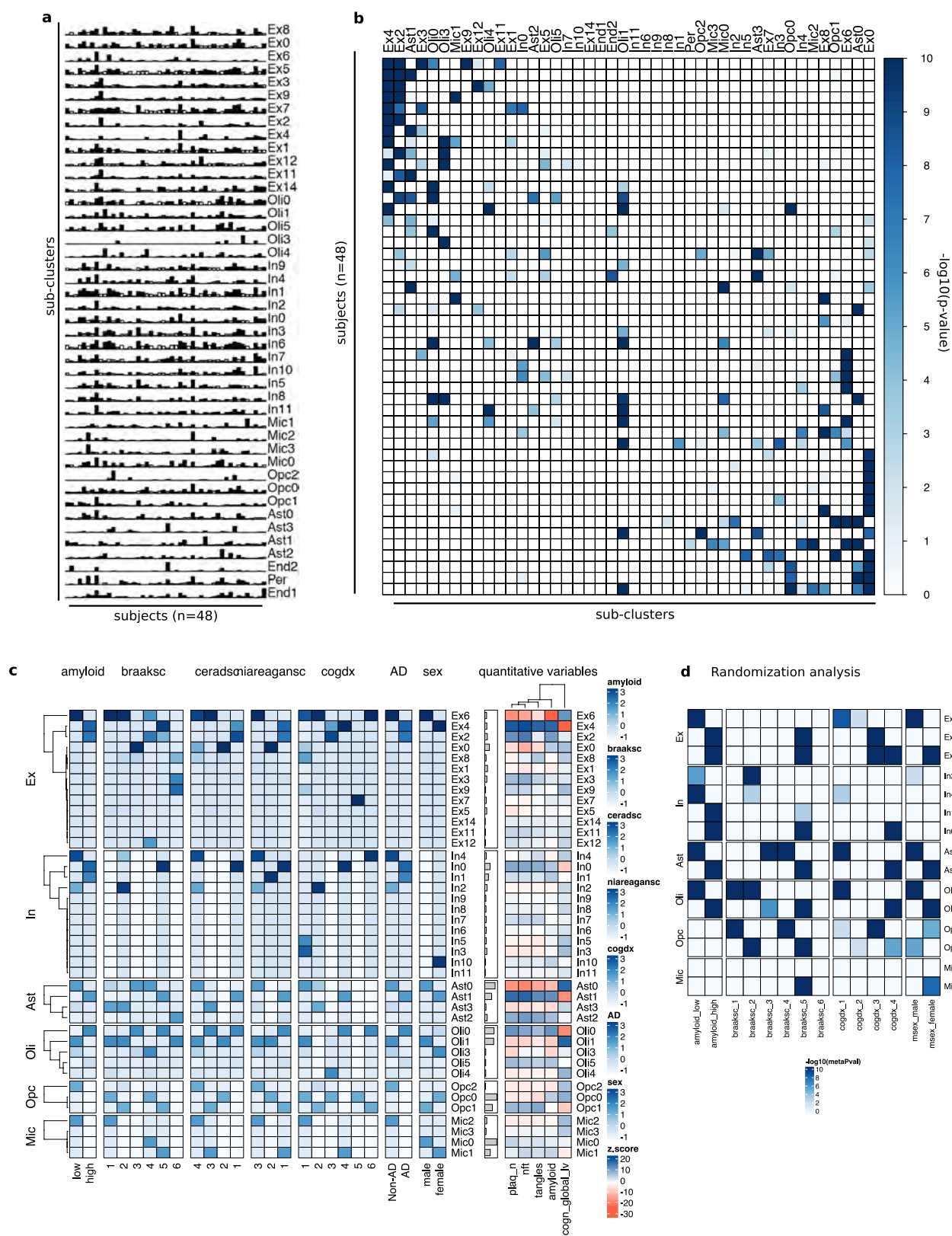
Extended Data Fig. 5 | Overlap of genes that are altered in the progression of AD pathology. Quantification of the overlap (Jaccard coefficient) between pairs of gene sets identified as differentially expressed

in each of the major cell types when comparing cells isolated from AD-pathology individuals with cells isolated from no-pathology individuals, and combinations of early- and late-pathology individuals.



Extended Data Fig. 6 | Cell-type-specific and phenotype-specific gene–trait correlation analysis. **a–e**, SOM generated from transcriptome-wide gene-expression correlation of each gene with neuropathological signatures of AD. Genes with similar correlation patterns are mapped to the same SOM unit and similar units group close together. SOM grid layout is common and built jointly across all phenotypes and all cell types. Colour indicates the average Spearman’s rank correlation for genes in each

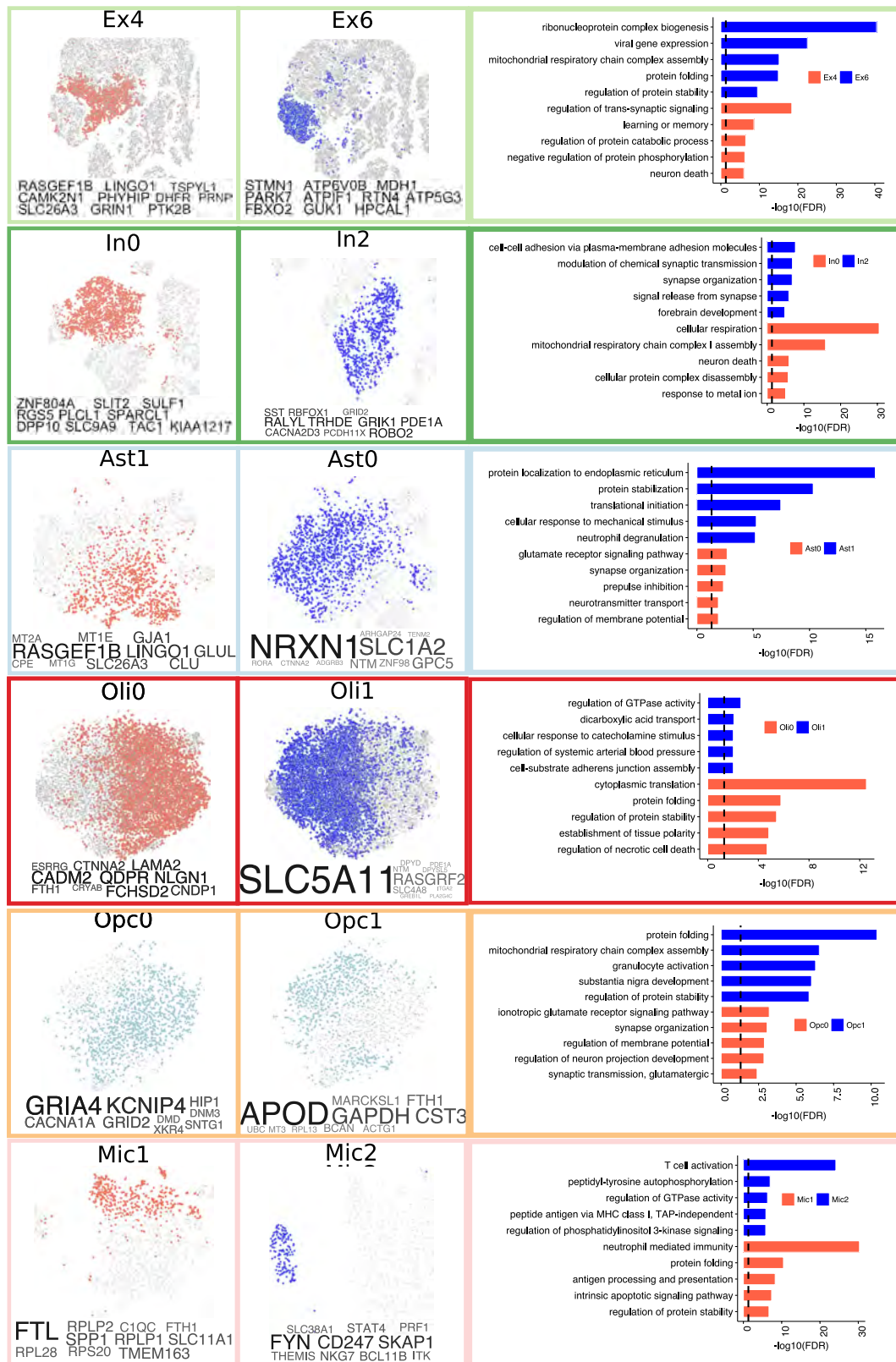
unit. **f**, Selected SOM territories (M1–M10). **g**, Overlap (one-sided Fisher’s exact test) between gene-trait correlation-module genes ($n = 1,472$ genes (M3), $n = 70$ (M6), $n = 80$ (M7)) and AD GWAS-risk genes (top; $n = 28$ genes), as well as genes associated with general cognitive function (bottom; $n = 709$ genes). The P values have been adjusted for multiple hypothesis testing; $-\log_{10}(\text{Bonferroni-corrected } P \text{ values})$ are shown.



Extended Data Fig. 7 | See next page for caption.

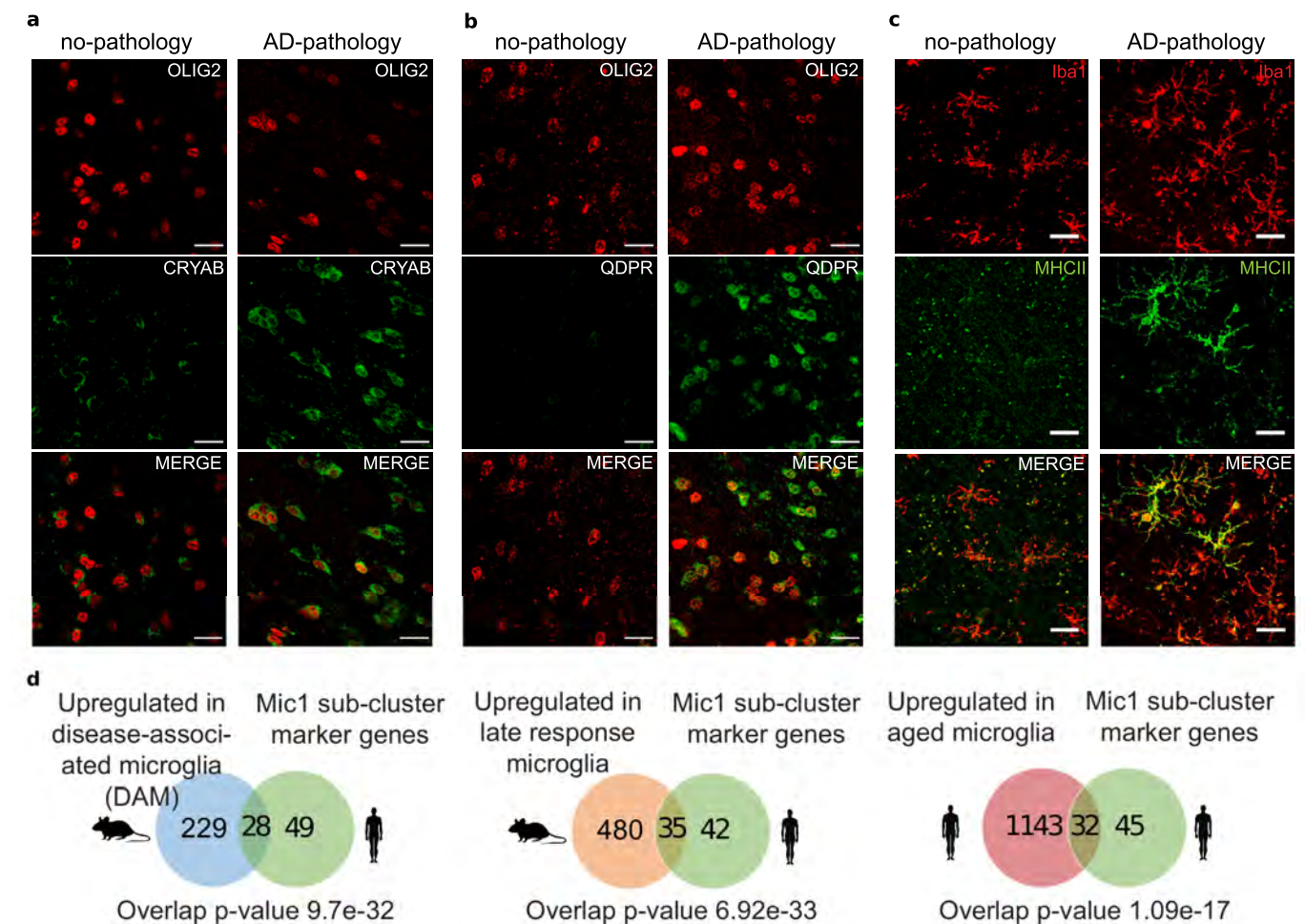
Extended Data Fig. 7 | Overrepresentation analyses for cells in sub-clusters. **a**, Cell composition of each identified sub-cluster (rows) across individuals (columns). Bars represent the fraction of cells corresponding to each individual. Bar colour indicates whether the corresponding value exceeds (black) or does not exceed (white) the average value measured across all the entries in the row. **b**, Overrepresentation analysis (hypergeometric test) within each pre-cluster (columns) of cells isolated from each individual (rows). **c**, Overrepresentation analysis within each sub-cluster of cells isolated from individuals with different values of discrete clinico-pathological variables (overall amyloid level, Braak stage, CERAD score (ceradsc), NIA-Reagan score (niareagansc), clinical consensus diagnosis of cognitive status at the time of death (cogdx), and sex). The scale bars on the right indicate the significance of the overrepresentation (hypergeometric test, $-\log_{10}(P \text{ value})$, z-scaled, FDR

multiple-testing correction). For quantitative variables, enrichment was computed based on an estimated z-score quantifying the deviation from random expected values using resampling (Methods). The quantitative variables considered were neuritic plaque count, neurofibrillary tangle burden, tangle density, overall amyloid level, and global cognitive function. For a detailed description of clinico-pathological variables, see Supplementary Information. **d**, Overrepresentation analysis (hypergeometric test) similar to that in **a**, but computed only across cells isolated from randomly chosen female and male individuals for AD-pathology and no-pathology groups (Methods). Scores represent aggregated P values (meta-p values, meanp method, metap R package) computed across 100 random realizations. Only scores with a FDR < 0.01 (correction across traits \times subpopulations) are plotted.



Extended Data Fig. 8 | Cell-type subpopulations. Cells from sub-clusters enriched (red) or depleted (blue) with cells for individuals with AD pathology and cognitive decline shown using *t*-SNE for major cell types (Ast1 $n = 1,134$, Ast0 $n = 1,728$, Oli0 $n = 8,310$, Oli1 $n = 8,032$, Ex4 $n = 3,198$, Ex6 $n = 2,757$, In0 $n = 2,368$, In2 $n = 984$, Opc0 $n = 1,589$, Opc1 $n = 976$, Mic1 $n = 509$, and Mic2 $n = 169$ cells). Left, corresponding marker genes (font proportional to enrichment level); right, enriched Gene

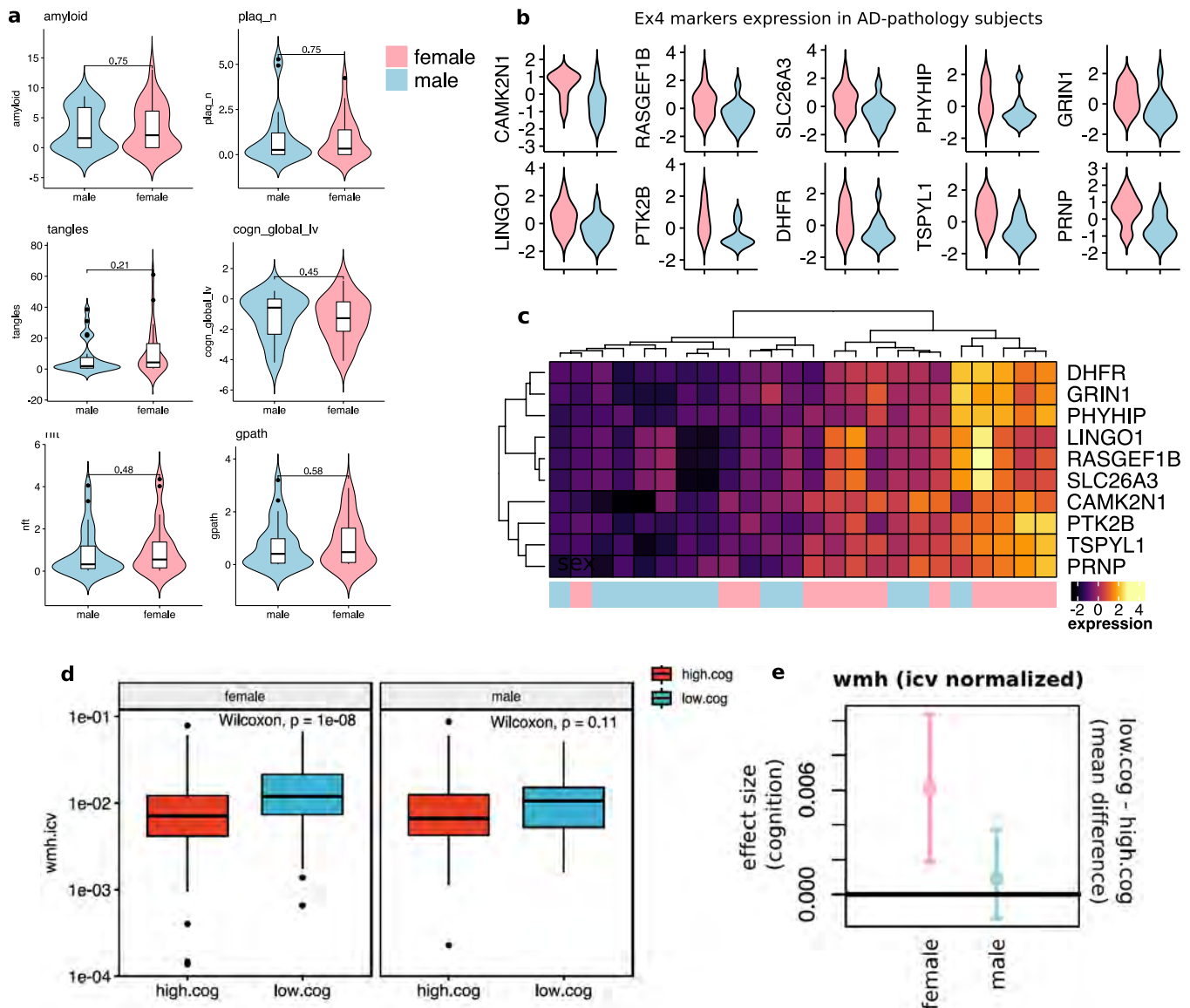
Ontology terms. Gene Ontology enrichment was based on FDR-corrected cumulative hypergeometric P values, with P value-ranked gene-marker lists ($FDR < 0.01$, $\log_2(\text{mean gene expression across cells in sub-cluster} / \text{mean gene expression across cells in other sub-clusters}) > 0.5$, two-sided Wilcoxon rank-sum test) used as input (Ex4 $n = 783$, Ex6 $n = 2,438$, In0 $n = 1,702$, In2 $n = 350$, Ast1 $n = 574$, Ast0 $n = 73$, Oli0 $n = 227$, Oli1 $n = 73$, Opc0 $n = 19$, Opc1 $n = 536$, Mic1 $n = 487$, and Mic2 $n = 646$).



Extended Data Fig. 9 | Immunohistochemistry of subpopulation markers in oligodendrocyte lineage cells and microglia.

a, Oligodendrocyte lineage cell subpopulation marked by alpha B-crystallin (CRYAB). Immunohistochemistry with anti-OLIG2 (red) and anti-CRYAB (green) antibodies in the white matter of Brodmann area 10 of a no-pathology and an AD-pathology individual (scale bars, 20 μm). A selected area of these images is shown in Fig. 3g. The experiment was performed once. **b**, Oligodendrocyte lineage cell subpopulation marked by quinoid dihydropteridine reductase (QDPR). Immunohistochemistry with anti-OLIG2 (red) and anti-QDPR (green) antibodies in the white matter

of Brodmann area 10 of a no-pathology and an AD-pathology individual (scale bars, 20 μm). A selected area of these images is shown in Fig. 3h. The experiment was performed once. **c**, Immunohistochemistry with anti-IBA1 (red) and anti-MHC class II (green) antibodies in the white matter of Brodmann area 10 of a no-pathology and an AD-pathology individual (scale bars, 20 μm). The experiment was performed once. **d**, Overlap (one-sided Fisher's exact test) between Mic1 marker genes and genes upregulated in mouse disease-associated microglia (left), in mouse late-response microglia (middle), and in aged human microglia (right).



Extended Data Fig. 10 | Sex comparisons in pathology, gene expression, and white matter. **a**, Quantitative clinico-pathological measurement comparison between male and female individuals ($n = 24$ female and $n = 24$ male individuals; two-sided Wilcoxon rank-sum test). Violin plots are centred around the median with interquartile ranges, and the shape represents individual distribution. The quantitative clinico-pathological variables considered were overall amyloid level, neuritic plaque burden, neurofibrillary tangle burden, tangle density, global cognitive function, and global AD pathology burden. **b**, Violin plots showing aggregate expression levels (z-scaled) across excitatory neurons in female (red) versus male (blue) individuals ($n = 12$ each) of the top 10 marker genes of the AD-associated Ex4 subpopulation of excitatory neurons. **c**, Hierarchical clustering of pathology-affected individuals (columns) based on average expression level (colour) of the top 10 marker genes (rows) of the AD-enriched Ex4 subpopulation of excitatory neurons for female versus male individuals. **d**, **e**, Statistical comparison of in vivo brain MRI imaging from ROSMAP cohorts. **d**, Intracranial volume-normalized WMH (wmh.icv) measures for female ($n = 399$) and male

($n = 106$) individuals and high-cognition ($n = 252$ female and $n = 63$ male individuals) and low-cognition ($n = 147$ female and $n = 43$ male individuals) groups. Groups were defined based on whether subjects had an overall cognition score lower (low.cog, z-score < 0) or higher (high.cog, z-score > 0) than the average. Mean rank-difference values between cognition groups were compared using the two-sided Wilcoxon rank-sum test. **e**, Statistical estimation of significant difference in WMH between low-cognition and high-cognition groups in females, and between low-cognition and high-cognition groups in males, assessed by bootstrap point and 95% confidence interval estimation of the effect size (mean difference) between groups. Bootstrap resampling was performed by resampling $n = 40$ observations per group 1,000 times. Horizontal line highlights zero difference. The positive effect-size points and confidence interval estimates do not overlap the zero line in the female group, which provides statistical evidence of an increment in WMH (wmh.icv) in the low-cognition group relative to the high-cognition group in females but not in males.

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

- | n/a | Confirmed |
|-------------------------------------|--|
| <input type="checkbox"/> | <input checked="" type="checkbox"/> The exact sample size (<i>n</i>) for each experimental group/condition, given as a discrete number and unit of measurement |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> The statistical test(s) used AND whether they are one- or two-sided
<i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i> |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> A description of all covariates tested |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> For null hypothesis testing, the test statistic (e.g. <i>F</i> , <i>t</i> , <i>r</i>) with confidence intervals, effect sizes, degrees of freedom and <i>P</i> value noted
<i>Give P values as exact values whenever suitable.</i> |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> Estimates of effect sizes (e.g. Cohen's <i>d</i> , Pearson's <i>r</i>), indicating how they were calculated |

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection

no software was used

Data analysis

Cell Ranger software (2.0.0 version) (10x Genomics) was used to align clean reads to the hg38 human genome (GRCh38.p5 (NCBI:GCA_000001405.20)). The single-cell analysis package scanpy (version 0.4.4) implemented in python was used for clustering analyses. The R packages scran ((version 1.8.2) and scater ((version 1.8.1) were used for single-cell data manipulations and QC analyses. The R package Seurat (2.3.4) was used to compared reproducibility of markers. Bulk RNA-seq differential analyses we performed by fitting a linear model using the R package limma (version 3.38.3). The R package lme4 (1.1-21) was used to fit Poisson mixed models, the R package RUVseq (1.16.1) to remove unwanted variation from RNA data, and the R package metap (1.1) to compute aggregate p-values. All statistical analyses and visualizations were implemented in R (version 3.4.3). GSEA was applied to identify priori defined gene sets that show statistically significant differences between two given clusters. Gene Ontology (GO) enrichment analyses were performed using Metascape (version 3.0). Raw MPAGE images were processed to generate total volumes including gray matter, white matter, CSF and intracranial volumes using SPM (version 12). White matter lesions appearing hyperintense in T2-weighted images were segmented based on FLAIR and MPAGE data using BIANCA. All self-organizing maps were created using the Kohonen R package (version 3.0.8).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The single-nucleus RNA-Sequencing data is available at Synapse (<https://www.synapse.org/#!Synapse:syn18485175>). The DOI for this dataset is: 10.7303/syn18485175. The data is available under controlled use conditions set by human privacy regulations.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No explicit calculations were performed to determine sample size. Rather, we aimed to analyze brain tissue from an equal number of men and women and at least 12 individuals per group. Therefore we analyzed brain tissue from 24 individuals with low amyloid burden and 24 individuals with high amyloid burden. These numbers of samples was sufficient to perform a confident data analysis.
Data exclusions	Low quality snRNA-seq libraries were excluded and the exclusion criteria are described in the manuscript as follows. Cells with a high ratio of Mitochondrial (MT) relative to endogenous RNAs had low starting amounts of RNA, which might indicate that source cells were dead or stressed, resulting RNA degradation. Outlier cells in these quality metrics were found to cluster together in the tSNE 2D space. Based on these observations and subsequent scatter plot analyses, cells with less than 200 detected genes, and cells with abnormally high ratio of counts mapping to MT genes, relative to the total number of detected genes were removed. Specifically, given a highly skewed empirical distribution of the MT ratio values (i.e., having an elbow shape clearly separating high and low scores) outlier cells were classified in two groups using the k-means clustering algorithm (k=2) on the MT ratio, and subsequently removed.
Replication	Verification of the single-nucleus RNA-seq data was performed through validation using RNA in situ hybridization, quantitative RT-PCR, and immunohistochemistry on tissue derived from a subset of the individuals analyzed using snRNA-seq. These experiments validated the findings derived from snRNA-seq.
Randomization	The study participants were allocated into groups based on the overall amyloid level.
Blinding	Investigators were blinded to group allocation for the quantification of the RNAscope data.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
<input type="checkbox"/>	<input checked="" type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input type="checkbox"/>	<input checked="" type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data

Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input type="checkbox"/>	<input checked="" type="checkbox"/> MRI-based neuroimaging

Antibodies

Antibodies used

anti-Iba1 (Synaptic Systems; Cat. No. 234 004, Polyclonal Guinea pig antiserum, Lot number: 2-13, 1:500);
anti-Human HLA-DP, DQ, DR antigen (Agilent, M077501-2, clone CR3/43, Lot number: 20047190, 1:100);
anti-β-Amyloid (Cell Signaling Technology, #8243, D54D2, Lot number: 4, 1:500);
anti-OLIG2 (Atlas, HPA003254, rabbit polyclonal, Lot number: C114365, 1:1000);

anti-CRYAB (LSBio, LS-B3696, rabbit polyclonal, Lot number: 124639, 1:200);
anti-QDPR (Atlas Antibodies, HPA065649, rabbit polyclonal, Lot number: R92923, 1:2500);
anti-Rabbit IgG (goat), HRP-labeled (PerkinElmer, NEF812001EA, from goat serum, Lot number: 10311573, Dilution);
Alexa Fluor®488 conjugated anti-NeuN antibody (MilliporeSigma, catalog number MAB377X, clone A60, Lot number: 3101114, 1:500)

Validation

anti-Iba1: reactivity validated by the company for Human. Validated by the company for IHC.
anti-Human HLA-DP, DQ, DR antigen: reactivity validated by the company for Human. The antibody was included in the First International Workshop and Conference on Monoclonal Antibodies to Human MHC Class II Antigens (1983) and its specificity and other characteristics were ascertained by a variety of techniques, including reactivity with isolated antigen, immunoblotting, and labelling of transfected cells.
anti-β-Amyloid: reactivity validated by the company for Human. Validated by the company for IHC on paraffin-embedded human Alzheimer's brain tissue sections.
anti-OLIG2: reactivity validated by the company for Human. Validated by the company for IHC. Has been validated by the Human Protein Atlas in 44 human control brain samples.
anti-CRYAB: reactivity validated by the company for Human. Validated by the company for IHC.
anti-QDPR: reactivity validated by the company for Human. Validated by the company for IHC. Has been validated by the Human Protein Atlas in 44 human control brain samples.
anti-Rabbit IgG (goat), HRP-labeled: Tested by the company to react with rabbit IgG and may recognize other immunoglobulin types that have light chains in common with IgG.
Alexa Fluor®488 conjugated anti-NeuN: reactivity validated by the company for Human.

Human research participants

Policy information about [studies involving human research participants](#)

Population characteristics

We selected 24 individuals with elevated β-amyloid pathology and 24 age-matched control individuals with no or very low β-amyloid burden. Individuals were balanced between male and female subjects (12 in each group), and matched for both age (median 86.7 high-amyloid, 87.1 low-amyloid) and years of education (median 19.5 high-amyloid, 18 low-amyloid).

Recruitment

No donors were recruited, the tissue has been obtained from participants in the Religious Order Study.

Ethics oversight

The Religious Orders Study and Rush Memory and Aging Project were approved by an IRB of Rush University Medical Center.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Magnetic resonance imaging

Experimental design

Design type

N/A

Design specifications

N/A

Behavioral performance measures

N/A

Acquisition

Imaging type(s)

Structural

Field strength

3T

Sequence & imaging parameters

3D magnetization-prepared rapid acquisition gradient echo (MPRAGE) sequence: echo-time (TE)=2.98 ms, repetition time (TR)=2.3 s, inversion time (TI)=900 ms, flip angle=8 degrees, 176 sagittal slices, slice thickness=1 mm, field of view (FOV)=25.6 cm x 25.6 cm, 256x256 acquisition matrix. T2-weighted fluid-attenuated inversion recovery (FLAIR): TE=150 ms, TR=9 s, TI=2.49 s, 35 axial slices, slice thickness=4 mm, FOV=22 cm x 22 cm, 256x256 acquisition matrix.

Area of acquisition

Whole brain

Diffusion MRI

☐ Used ☒ Not used

Preprocessing

Preprocessing software

FLIRT (affine registration of T1-weighted MPRAGE to T2-weighted FLAIR), BET (brain extraction), WMLS (automated white matter hyperintensity segmentation based on both T1-weighted MPRAGE and T2-weighted FLAIR), Freesurfer (intracranial volume calculation).

Normalization

T1-weighted MPRAGE data were spatially registered to the T2-weighted FLAIR data using affine registration.

Normalization template

The data were not normalized to a standardized template.

Noise and artifact removal

N/A

Volume censoring

N/A

Statistical modeling & inference

Model type and settings

The total volume of white matter hyperintensities was measured for each participant and then normalized by the corresponding intracranial volume.

Effect(s) tested

N/A

Specify type of analysis: ☒ Whole brain ☐ ROI-based ☐ BothStatistic type for inference
(See [Eklund et al. 2016](#))

N/A

Correction

N/A

Models & analysis

n/a | Involved in the study

☒ ☐ Functional and/or effective connectivity☒ ☐ Graph analysis☒ ☐ Multivariate modeling or predictive analysis

Structural mechanism for NEK7-licensed activation of NLRP3 inflammasome

Humayun Sharif^{1,2,9}, Li Wang^{1,2,9}, Wei Li Wang^{1,2,3,4,5,6,7,9}, Venkat Giri Magupalli^{1,2}, Liudmila Andreeva^{1,2}, Qi Qiao^{1,2}, Arthur V. Hauenstein^{1,2}, Zhaolong Wu^{3,4}, Gabriel Núñez⁸, Youdong Mao^{3,4,5,6,7*} & Hao Wu^{1,2*}

The NLRP3 inflammasome can be activated by stimuli that include nigericin, uric acid crystals, amyloid- β fibrils and extracellular ATP. The mitotic kinase NEK7 licenses the assembly and activation of the NLRP3 inflammasome in interphase. Here we report a cryo-electron microscopy structure of inactive human NLRP3 in complex with NEK7, at a resolution of 3.8 Å. The earring-shaped NLRP3 consists of curved leucine-rich-repeat and globular NACHT domains, and the C-terminal lobe of NEK7 nestles against both NLRP3 domains. Structural recognition between NLRP3 and NEK7 is confirmed by mutagenesis both in vitro and in cells. Modelling of an active NLRP3-NEK7 conformation based on the NLRP4 inflammasome predicts an additional contact between an NLRP3-bound NEK7 and a neighbouring NLRP3. Mutations to this interface abolish the ability of NEK7 or NLRP3 to rescue NLRP3 activation in NEK7-knockout or NLRP3-knockout cells. These data suggest that NEK7 bridges adjacent NLRP3 subunits with bipartite interactions to mediate the activation of the NLRP3 inflammasome.

Inflammasomes are cytoplasmic supramolecular complexes that form in response to exogenous microbial invasions and endogenous damage signals^{1–3}. Inflammasomes activate inflammatory caspases such as caspase-1, which processes the proinflammatory cytokines interleukin 1 β (IL-1 β) and IL-18 for their maturation, and cleaves gasdermin D to generate the N-terminal fragment to induce pore formation, cytokine release and pyroptotic cell death^{2,4,5}. NLRP3 belongs to the family of nucleotide-binding domain (NBD) and leucine-rich repeat (LRR)-containing proteins (NLRs). It has an N-terminal pyrin domain, which interacts with the adaptor protein ASC via interactions between pyrin domains; a central adenosine triphosphatase (ATPase) domain known as NACHT, which comprises an NBD, helical domain 1 (HD1), winged helix domain (WHD) and helical domain 2 (HD2); and a C-terminal LRR domain⁶ (Fig. 1a). ASC also has a caspase recruitment domain, which recruits caspase-1 via interactions between the caspase recruitment domains, to promote caspase dimerization and activation. Previous studies have revealed that the pyrin and caspase recruitment domains both form filamentous assemblies through nucleated polymerization^{7–9}. NLRP3 is an extensively studied inflammasome sensor activated by a spectrum of seemingly unrelated stimuli via induction of K⁺ efflux^{1–3,10–12}. Autosomal-dominant mutations of the *NLRP3* gene are related to autoinflammatory diseases that are collectively known as cryopyrin-associated periodic syndromes (CAPS); NLRP3 hyperactivation is directly connected to systematic and joint inflammation^{1–3,10}. The mitotic Ser/Thr kinase NEK7 has recently been identified as an important requirement in NLRP3 activation via direct NLRP3-NEK7 interaction^{13–16}. Because of its interaction with NEK9 during mitosis, and its limited quantity in cells, NEK7 licenses NLRP3 activation only in interphase¹⁶. However, the molecular mechanism of the NLRP3-NEK7 interaction remains unknown.

Cryo-electron microscopy structure determination

We expressed and purified a recombinant complex of maltose-binding protein (MBP)-tagged NLRP3 with the pyrin domains deleted (NLRP3 Δ) and NEK7. Microscale thermophoresis showed a dissociation constant of 78.9 ± 38.5 nM between NLRP3 and NEK7 (Fig. 1b, Extended Data Fig. 1a). A thermal shift assay indicated that the NLRP3 inhibitor MCC950¹⁷ and ADP both increased the stability of NLRP3 or of the complex, but not of NEK7 alone (Extended Data Fig. 1b–d). However, cryo-electron microscopy (cryo-EM) images of the complex showed poor contrast, probably owing to the relatively small size of the complex (about 185 kDa with the MBP tag).

To facilitate structure determination, we generated an artificial NEK7 dimer that was based on a dimer structure of protein kinase R (RCSB Protein Data Bank (PDB) code 2A19)¹⁸ (Extended Data Fig. 1e). The engineered NEK7 formed a dimeric complex with NLRP3, as shown by a shift in the elution position from a gel filtration column and by multi-angle light scattering measurements (Fig. 1c, d, Extended Data Fig. 1f). Although dimeric in solution, 2D classification revealed mostly monomeric NLRP3-NEK7 particles, which suggests that the two complexes in the dimer are flexibly linked. We speculate that, despite the flexibility, the larger dimer complexes facilitated particle-picking and 3D reconstruction. Cryo-EM analysis was first conducted using a Talos Arctica microscope, which resulted in a 4.3 Å reconstruction (Extended Data Fig. 2a–d, Extended Data Table 1). A larger cryo-EM dataset that contained 15,681 movies collected on a Titan Krios microscope (Extended Data Fig. 2e) gave rise to the final 3.8 Å map, refined with 108,771 particles selected from multiple rounds of 2D and 3D classifications (Fig. 1e, f, Extended Data Fig. 2e–g, Extended Data Table 1). The data included 1,340 movies collected at a stage tilt of 20°, which improved particle orientation distribution (Extended Data Fig. 2f).

Atomic model building was initially performed on the 4.3 Å map, assisted by structures of other NLRs and NEK7, and was finalized on

¹Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, MA, USA. ²Program in Cellular and Molecular Medicine, Boston Children's Hospital, Boston, MA, USA. ³State Key Laboratory for Artificial Microstructures and Mesoscopic Physics, School of Physics, Peking University, Beijing, China. ⁴Center for Quantitative Biology, Peking University, Beijing, China. ⁵Intel Parallel Computing Center for Structural Biology, Dana-Farber Cancer Institute, Boston, MA, USA. ⁶Department of Cancer Immunology and Virology, Dana-Farber Cancer Institute, Boston, MA, USA. ⁷Department of Microbiology, Harvard Medical School, Boston, MA, USA. ⁸Department of Pathology and Comprehensive Cancer Center, University of Michigan Medical School, Ann Arbor, MI, USA. ⁹These authors contributed equally: Humayun Sharif, Li Wang, Wei Li Wang. *e-mail: ymao@pku.edu.cn; wu@crystal.harvard.edu

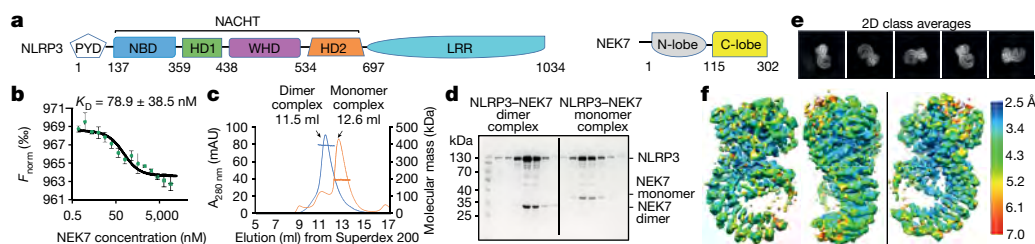


Fig. 1 | Biochemical characterization and cryo-EM structure determination. **a**, Schematic domain representation of human NLRP3 and NEK7, with labelled domain boundary defined from this work. PYD, pyrin domain. **b**, Microscale thermophoresis analysis of NEK7 binding to NLRP3. A dissociation constant of 78.9 ± 38.5 nM was calculated from three independent replicates (shown as mean \pm s.d.). **c**, Reconstitution of NLRP3-NEK7 dimer and monomer complexes on Superdex 200 gel filtration column (repeated ≥ 5 times). Molecular mass

distribution within each peak was calculated from in-line multi-angle light scattering measurements, and is shown in blue and orange for the dimer and monomer complexes, respectively (performed once). mAU, milli-absorbance units. **d**, SDS-PAGE gels of eluted fractions of dimer and monomer complexes (repeated ≥ 5 times). **e**, Representative 2D class averages from the 300 keV cryo-EM dataset, selected amongst 100 classes. **f**, The final cryo-EM density map (shown at 6σ) in three orientations and coloured with local resolutions by ResMap⁴¹.

the 3.8 Å map (Fig. 2a, Extended Data Fig. 3). The map revealed an earring shape—which is characteristic of NLRs, such as NLRC4^{6,19,20}, NOD2²¹ and NAIP5^{22,23}—that contains a curved LRR domain and a compact NACHT that comprises an NBD, HD1, WHD and HD2 (Fig. 2a). The LRR region comprises 12 repeats, consistent with the structure prediction using Phyre2²⁴ (Extended Data Fig. 4). The density encircled by the LRR fit well with the C-terminal lobe of the crystal structure of human NEK7 (PDB code 2WQN)²⁵ as a rigid body. No density was visible for the N-terminal lobe of NEK7 (Extended Data Fig. 3a, h), probably because the mutations introduced to this region to generate a NEK7 dimer affected the structural integrity of NEK7 N-lobe.

Overall structure of the complex, and interactions

The final atomic model contains almost the entire NLRP3 Δ (without the N-terminal MBP), the NEK7 C-lobe and the ADP bound to the NBD of NLRP3 (Fig. 2a, Extended Data Table 1, Supplementary Video 1). When the full-length NEK7 structure is overlaid with the NLRP3-NEK7 model, the N-lobe extends away from the complex with no steric hindrance against NLRP3 (Extended Data Fig. 5a). The activation loop, including S195 (the phosphorylation of which activates NEK7), is disordered in the NLRP3-NEK7 complex and does not contact NLRP3 (Extended Data Fig. 5b), which is consistent with the observation that both active and inactive NEK7 can support NLRP3 activation^{13–15}. The NBD-HD1-WHD module of NLRP3 exhibits a conformation that is similar to the inactive conformation of NLRC4 and NOD2, whereas the HD2 and LRR structures are variable among known structures in the NLR family (Extended Data Fig. 5c–e). NEK7 binds an LRR region in which the phosphorylated S533 residue of NLRC4 HD2 interacts intramolecularly, although the role of S533 phosphorylation in NLRC4 activation remains controversial^{6,26,27} (Extended Data Fig. 5e).

NLRP3 interacts with NEK7 at multiple surfaces, including the LRR, HD2 and NBD (Extended Data Figs. 6, 7). The calculated iso-electric points of NEK7 and NLRP3 are 8.5 and 6.2, respectively, which makes NEK7 positively charged overall, and NLRP3 negatively charged overall, at physiological pH (Fig. 2b). The NEK7 C-lobe is even more positively charged, with a calculated iso-electric point of 9.0. Indeed, the interaction between NLRP3 and NEK7 is dictated partly by electrostatic complementarity (Fig. 2b). The positively charged nature of NEK7 is also reflected in the bound SO_4 ion in its crystal structure²⁵, at its interface with NLRP3 (Extended Data Fig. 8a). The interactions between NLRP3 and NEK7 can be divided into two interfaces: those between the LRR of NLRP3 and the first half of the NEK7 C-lobe (Fig. 3a, b, Extended Data Fig. 8b) and those between NACHT (NBD and HD2) and the second half of the NEK7 C-lobe (Fig. 3a, c, Extended Data Fig. 8b). These interactions bury surface areas of about 880 Å² and about 770 Å², respectively.

Structural insights into NEK7 for NLRP3 interaction

On NEK7, residues with largest buried surface areas at the interface (as assessed by PISA²⁸) include Q129, R131 and R136; these residues interact with the LRR domain of NLRP3 (Fig. 3b). R121, R131 and R136 also participate in hydrogen bonding or charged interactions. The NEK7 residues D261, E265 and E266 bury large surface areas at the interface with HD2 of NLRP3 (Fig. 3c, left). S260 and E266 also participate in hydrogen bonding or charged interactions at this interface. For the NBD interaction, the NEK7 residues D290, K293 and R294 appear to be the most extensive (Fig. 3c, right).

To deduce the importance of the observed contacts, we performed site-directed mutagenesis on NEK7 (Fig. 3d, e, Extended Data Figs. 6, 8c–e). The pulldown of SUMO-tagged wild-type and mutant NEK7 by MBP-tagged NLRP3 using amylose resin revealed that the most-marked effects came from mutations on Q129, R131 and R136 in the first half of the NEK7 C-lobe, all at the interface with the LRR of NLRP3 (interface I). The data explain previous immunoprecipitation results, in which a truncated NEK7 (amino acids 1–212, with the second half of the C-lobe removed) supported NLRP3 binding¹³. The residues S260, D261 and E265 of NEK7 for the HD2 interaction (interface II) also had substantial effects when they were mutated, whereas mutations

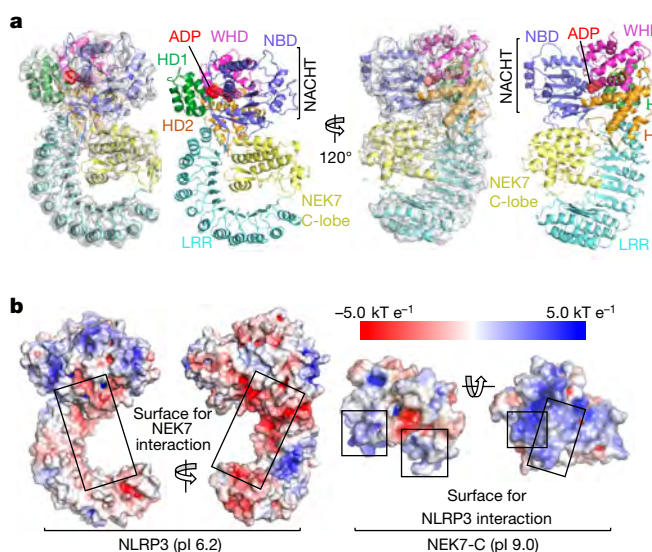


Fig. 2 | Cryo-EM structure overview. **a**, Ribbon diagrams of the complex, shown with and without the cryo-EM map (grey, 5σ) in two orientations. Domains are colour-coded as in Fig. 1a and the bound ADP is shown in sphere rendering. **b**, Electrostatic surface representation of NLRP3 and NEK7, colour-coded according to electrostatic potential from red (-5.0 kT e⁻¹, negatively charged) to blue (5.0 kT e⁻¹, positively charged). pI, iso-electric point.

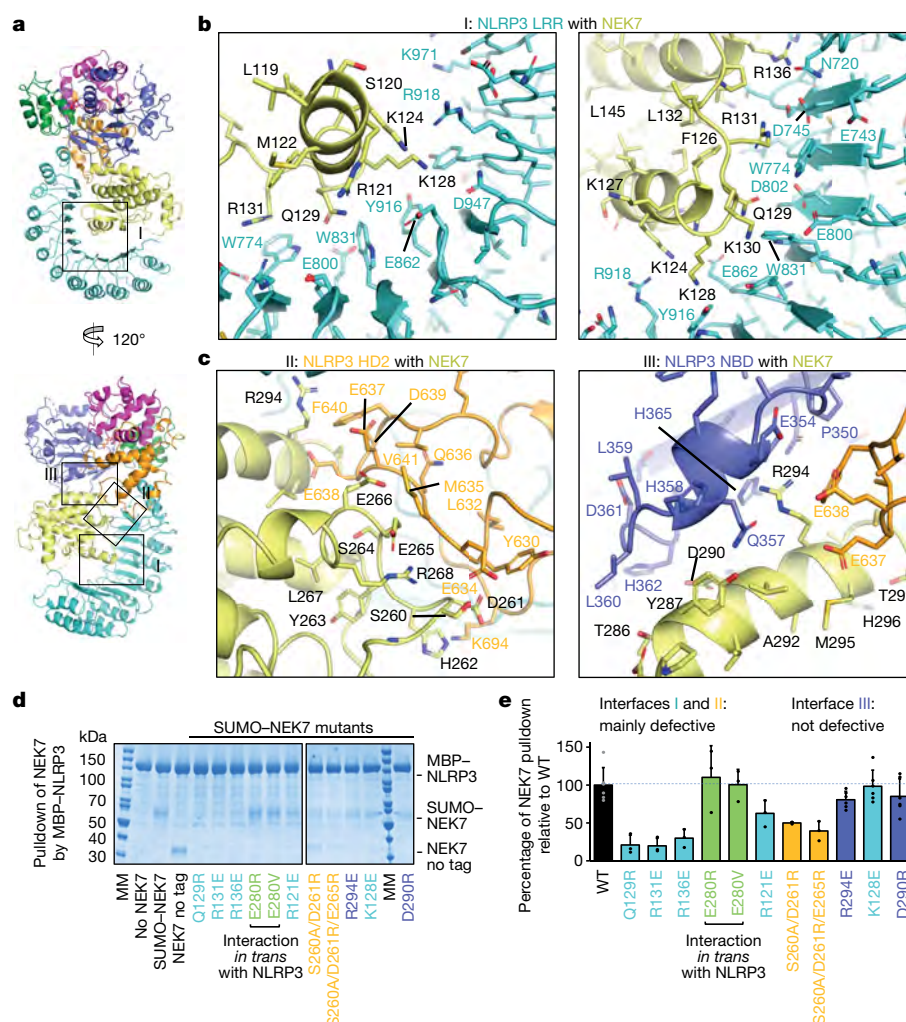


Fig. 3 | Structural insights into NEK7 interaction with NLRP3. **a**, The NLRP3–NEK7 complex in two orientations to denote the three interaction regions (I, II and III). **b**, **c**, Magnified views of interaction interfaces between NEK7 and LRR of NLRP3 (**b**) and between NEK7 and HD2 and NBD of NLRP3 (**c**), with main chains in cartoons and side chains in sticks. **d**, Pulldown of purified wild-type and mutant His-SUMO-tagged NEK7 by

purified MBP-tagged wild-type NLRP3 using amylose resin. Experiments were performed between three and six times. NEK7 mutants are coloured by the domain of NLRP3 that they contact, as in **a–c**. MM, molecular mass markers. **e**, Quantification of mutant NEK7 binding to NLRP3 relative to wild-type (WT) NEK7 (shown as mean \pm s.d., $n = 3–6$ replicates). Dots, individual data points. For gel source data, see Supplementary Fig. 1.

on NEK7 residues for the NBD interaction did not cause notable changes in NLRP3 interaction (Fig. 3d, e, Extended Data Figs. 6, 8c–e). Collectively, the mutagenesis data support the importance of interfaces I and II in the NLRP3–NEK7 interaction. Of note, the NLRP3 NBD no longer contacts NEK7 in a model of active NLRP3 based on the active NLRP4 conformation^{19,20} (see below).

To further validate the observed NLRP3–NEK7 interface, we reconstituted wild-type and mutant NEK7 into NEK7-knockout immortalized mouse bone-marrow-derived macrophages (iBMDMs) and examined the response of the cells upon lipopolysaccharide (LPS) priming and nigericin treatment. Whereas wild-type NEK7 rescued NEK7-knockout iBMDMs in caspase-1 processing (Fig. 4a), IL-1 β secretion (Fig. 4b) and cell death (Extended Data Fig. 8f), the NEK7 mutants were compromised in supporting nigericin-induced NLRP3 activation (Fig. 4a, b, Extended Data Fig. 8f).

Specificity and competition in NEK7–NLRP3 interaction

NEK6 is a kinase that is closely related to NEK7, with an overall 87% sequence identity in the kinase domain (Extended Data Fig. 6); however, it cannot support NLRP3 activation^{13,14}. Sequence comparison among NEK proteins suggests that the main determinant of the specificity resides in the second part (residues 260–302) of the NEK7 C-lobe; the first part (residues 120–259) of the NEK7 C-lobe is largely

conserved with NEK6 (Extended Data Fig. 6). Because both NEK proteins are involved in mitosis through an interaction with NEK9, we mapped the NEK9-binding site onto the NEK7 structure from the crystal structure of a NEK7–NEK9 complex²⁹ (Extended Data Fig. 8g). Consistent with the sequence conservation, the NEK9-binding site is mainly formed by the first part of the NEK7 C-lobe (Extended Data Fig. 8g), and partially overlaps with the NLRP3-binding site of NEK7 (Extended Data Fig. 8h). In addition, the back-to-back homodimer of NEK7 in the NEK7–NEK9 crystal structure²⁹ that is postulated for *trans*-autophosphorylation is incompatible with NLRP3 interaction (Extended Data Fig. 8i). We hypothesize that, once NEK7 binds NLRP3, it can no longer interact with NEK9 for mitosis, and vice versa. Therefore, it is not only the case that mitosis prevents NLRP3 inflammasome activation¹⁴—NLRP3 inflammasome activation may also prevent mitosis, when the quantity of NEK7 in macrophages is limited¹⁴.

Structural insights into NLRP3 for NEK7 interaction

To test the functional relevance of NLRP3 residues that contact NEK7, we reconstituted wild-type and structure-based mutant NLRP3 into NLRP3-knockout iBMDMs and examined the response to NLRP3 stimulation. Upon LPS priming and nigericin treatment, wild-type reconstitution of NLRP3-knockout iBMDMs rescued the processing of caspase-1, as revealed by the appearance of the p20 large subunit

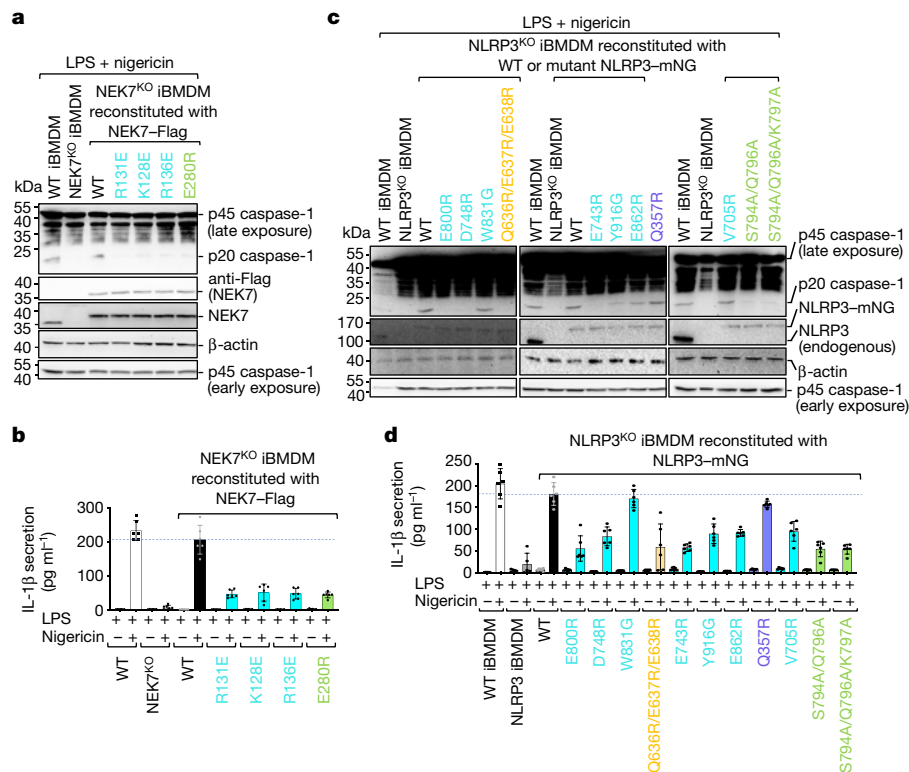


Fig. 4 | Structure-guided mutations of NLRP3 and NEK7 on inflammasome activation. NEK7 and NLRP3 mutations are coloured as in Fig. 3d, e, except that mutations on interactions *in trans* are coloured in green. **a**, **b**, NEK7-knockout (NEK7^{ko}) iBMDMs were reconstituted with wild-type or mutant Flag-tagged human NEK7, primed by LPS (4 h) and stimulated by nigericin (30 min). Cells were analysed by western blot for caspase-1 processing using a specific anti-caspase-1 antibody (repeated three times) (**a**). The full-length caspase-1 (p45) and processed large subunit of caspase-1 (p20) are labelled (first blot from the top). Caspase-1 p45 was also analysed by an early exposure to show equal loading (fifth blot). Reconstituted NEK7 was probed by anti-NEK7 and anti-Flag

antibodies (second and third blots), and loading was analysed by an anti-β-actin antibody (fourth blot). Mature IL-1β released in the supernatant was measured by enzyme-linked immunosorbent assay (**b**). Data are presented as mean ± s.d. for *n* = 3 replicates each from 2 independent experiments. Dots, individual data points. **c**, **d**, NLRP3-knockout (NLRP3^{ko}) iBMDMs were reconstituted with wild-type or mutant human mNeonGreen (mNG)-tagged NLRP3, primed by LPS (4 h) and stimulated by nigericin (30 min). Caspase-1 processing (repeated 3 times) (**c**) and IL-1β release (shown as mean ± s.d. with *n* = 3 replicates each from 2 independent experiments) (**d**) were analysed as in **a**, **b**. Dots, individual data points. For gel source data, see Supplementary Fig. 1.

(Fig. 4c, Extended Data Fig. 7). The LRR mutants E800R, D748R and E743R were strongly defective in caspase-1 processing; it is possible that Y916G, E862R and V705R were partially defective. The LRR mutant W831A had no discernible effect. A triple mutation in the HD2 region (Q636R/E637R/E638R) was strongly defective. By contrast, the Q357R mutant in the NBD was not defective. Similarly, nigericin-induced IL-1β secretion and cell death showed a similar pattern of mutational effects (Fig. 4d, Extended Data Figs. 7, 8j), which supports the notion that the LRR and HD2 domains—but not the NBD—are imperative for the interaction between NLRP3 and NEK7.

The importance of the LRR may explain why a partial LRR construct of NLRP3 (amino acids 742–991) did not interact with NEK7¹³, whereas a more-complete LRR domain construct (amino acids 711–1033) interacted with NEK7—albeit with some affinity reduction compared to the interaction with full-length NLRP3¹⁴. Consistently, NLRP3 phosphorylation at Y859 has a negative regulatory role³⁰; Y859 is situated at the NEK7-binding site of LRR, the phosphorylation of which might cause steric hindrance and charge repulsion (Extended Data Fig. 8k). On the other hand, the tyrosine phosphatase PTPN22 (which dephosphorylates Y859) promotes NLRP3 activation; patients with inflammatory bowel disease who carry an autoimmunity-associated allele of *PTPN22* show increased IL-1β production³⁰. Although the NACHT domain also contacts NEK7, the ability of NACHT alone to bind NEK7 remains unresolved. In one study, an LRR-deletion mutant (amino acids 1–741) did not interact with NEK7¹³. In another study, however, other LRR-deletion mutants (amino acids 1–686, 1–695, 1–710, 1–720 and 1–731) interacted with NEK7 and reconstituted NLRP3^{-/-} macrophages for NLRP3 activation by multiple stimuli³¹. Constructs with shorter LRR

truncations (amino acids 1–850, 1–879, 1–907, 1–936, 1–965 and 1–996) were inactive in reconstituting the activity of NLRP3³¹.

Oligomeric assembly of the NLRP3–NEK7 complex

Previous studies on NLRC4 have shown that the NACHT domain undergoes a large rigid-body rotation at the HD1-to-WHD junction to open up the structure for oligomerization and activation^{6,19,20,22,23}. We therefore modelled a hypothetical NLRP3 structure in an active conformation using the NLRC4 oligomer structure as a homology reference, which generates an approximately 90° rotation of the NBD–HD1 module (Fig. 5a). This modelling exercise placed the NBD–HD1–WHD module for direct interaction in the NLRP3 ring, without steric hindrance (Fig. 5b). By contrast, the inactive conformation is not compatible with the modelled oligomeric structure (Fig. 5c). The modelling should not affect NEK7 binding because only the NBD–HD1 module is moved relative to NEK7 (Extended Data Fig. 8l, Supplementary Video 2) and the NBD contact is not essential for NEK7 interaction, as shown by mutagenesis (Fig. 3d, e, 4c, d, Extended Data Fig. 8j).

Modelling of the hypothetical NLRP3–NEK7 inflammasome disc based on the NLRC4 oligomer places NEK7 at the oligomerization interface (Fig. 5b, d, Extended Data Fig. 9a). In particular, residue E280 of the NEK7 C-lobe (opposite from its NLRP3-interacting surface) contacts the neighbouring NLRP3 at residues S794, Q796 and K797, with a predicted buried surface area of 150 Å² (Fig. 5d). To test the functional importance of these predicted interactions in the modelled NLRP3 oligomer, we reconstituted NEK7-knockout or NLRP3-knockout iBMDMs using mutants on this new interface. As predicted, E280R or E280V did not affect the interaction between

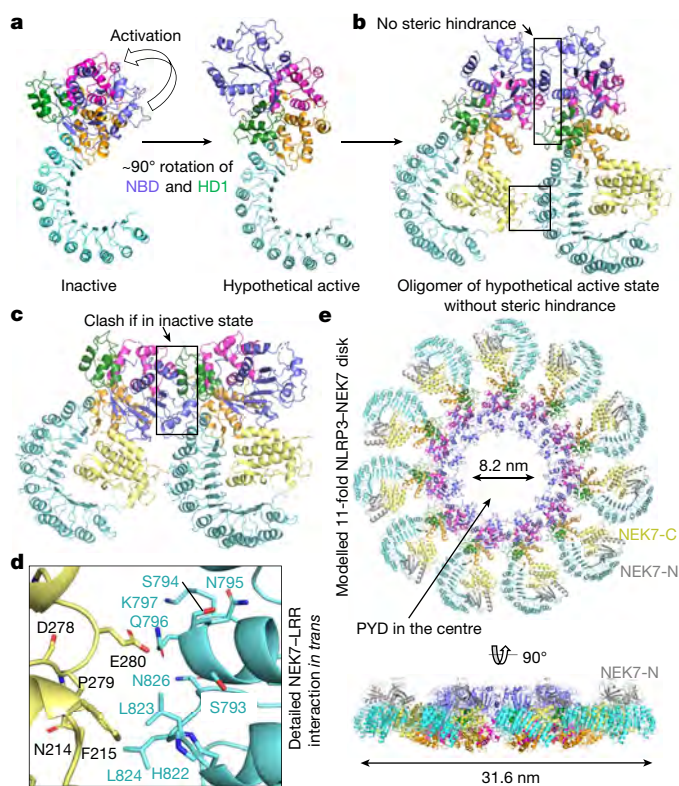


Fig. 5 | Modelling of active NLRP3 conformation and oligomerization.

a, Modelled active NLRP3 (right) from its inactive structure (left) using the activation mechanism of NLRC4, in which the NBD–HD1 module rotates by 90° relative to the WHD–HD2–LRR module^{19,20}. **b**, A dimer model of the active NLRP3–NEK7 complex based on the NLRC4 disk²⁰. Interaction surfaces are boxed. **c**, A hypothetical inactive NLRP3 dimer would have created steric clashes at multiple sites in the NACHT domain. **d**, Magnified view showing the detailed NEK7–LRR interaction *in trans* boxed in **b**. Mutations at this interface compromised NLRP3 inflammasome activation as in Fig. 4. **e**, NLRP3–NEK7 inflammasome disk modelled after the NLRC4 11-subunit disk structure²⁰, shown in two orientations.

NEK7 and NLRP3 in a pulldown assay using recombinant proteins, as E280 resides on the side of NEK7 opposite its NLRP3-binding site (Fig. 3d, e). However, NEK7(E280R) did not rescue NLRP3 activation in NEK7-knockout iBMDMs, whereas wild-type NEK7 did (Fig. 4a, b, Extended Data Fig. 8f). The double mutant NLRP3(S794A/Q796A) and the triple mutant NLRP3(S794A/Q796A/K797A) both failed to reconstitute NLRP3-induced caspase-1 processing, IL-1 β secretion and cell death in NLRP3-knockout iBMDMs, whereas wild-type NLRP3 did (Fig. 4c, d, Extended Data Fig. 8j). To confirm that NLRP3 inflammasome assembly *per se* was disrupted by these mutations, we assessed NLRP3 inflammasome speck formation using immunofluorescence imaging. Whereas wild-type NEK7 and NLRP3 rescued speck formation in NEK7-knockout or NLRP3-knockout iBMDMs, the NEK7 and NLRP3 mutants all failed to form speck in reconstituted iBMDMs upon nigericin stimulation (Extended Data Fig. 9b).

The NLRC4 inflammasome also has an additional oligomerization interface at the LRR (Extended Data Fig. 9c). Moreover, the recently identified NLRC4 autoinflammatory mutation W655C at this LRR–LRR interface (Extended Data Fig. 9d) promotes NLRC4 inflammasome assembly, probably by enhancing the LRR–LRR interaction, and causes macrophage activation syndrome in human³². In NLRP3, the LRR itself is too short to reach the adjacent LRR in the hypothetical oligomer, and NEK7 bridges the gap between adjacent NLRP3 subunits. Collectively, these data support our model of NLRP3 activation and point to a previously unanticipated mechanism for NEK7 requirement. In the modelled complex with full-length NEK7, the NEK7 N-lobe

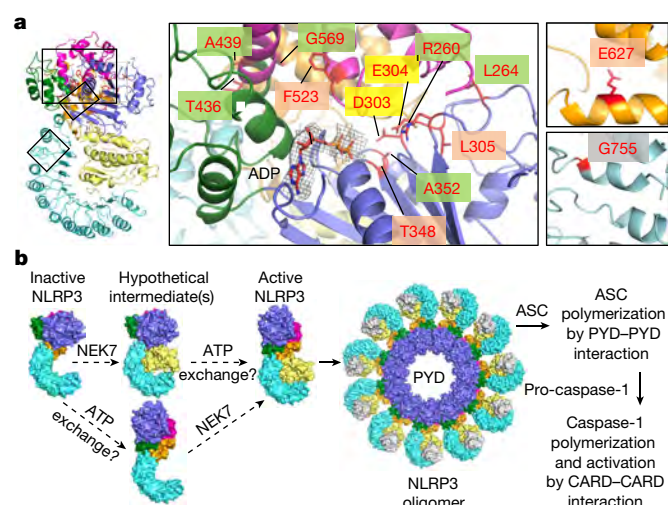


Fig. 6 | NLRP3 CAPS mutations and model of NEK7-mediated NLRP3 inflammasome activation. **a**, Structure mapping of all validated pathogenic mutations of NLRP3 from the Infervers website³³, highlighted by background colours behind residue labels. Density for ADP is shown (3 σ). Green, disruption of inter-domain interactions in the inactive conformation; mauve, change of local conformation by mutating residues buried within the domains; yellow, alteration in key residues in the Walker B motif; grey, enhancement of NEK7 binding. **b**, A proposed two-step NLRP3 inflammasome activation model. CARD, caspase recruitment domain.

projects away from the plane of the inflammasome ring and does not interfere with NLRP3 oligomerization (Fig. 5e).

NLRP3 CAPS mutations

To investigate the mechanism of autoinflammation in CAPS, we mapped the validated pathogenic mutations of NLRP3 from the Infervers website³³ to the structure (Fig. 6a, Extended Data Fig. 9e). These mutations are associated with Muckle–Wells syndrome, neonatal onset multisystem inflammatory disease and familial cold autoinflammatory syndrome. Notably, almost all the mutations lie in the NACHT domain and surround the ADP-binding site. One exception is G755 in the LRR near the NEK7-binding site, which exhibited enhanced interaction with NEK7 when mutated to alanine or arginine¹⁴.

Structural analysis suggests that the NACHT mutations may affect the inactive conformation, and therefore enhance NLRP3 activation. These mutations may be categorized into three tiers: (1) those that reside at the inter-domain interfaces, and thus affect the important interactions that stabilize the inactive conformation (R260, L264, D303, A352, T436, A439 and G569); (2) those that are on the Walker B motif of the ATPase (D303, which coordinates Mg²⁺ ions; and E304 for ATP hydrolysis³⁴); and (3) those that are buried within their respective domains (L305, T348, F523 and E627), in which mutations may affect the local domain conformation and indirectly destabilize the inter-domain interactions in the inactive state. Whereas L305, T348 and F523 are uncharged, E627 forms a partially buried salt-bridge with R548 and K568 in the NLRP3 structure.

Conclusion

In summary, our structural and functional studies have demonstrated that the NLRP3–NEK7 interaction dictates the NEK7 requirement for NLRP3 inflammasome activation. We further show that the opposite side of NEK7—which is not required for the initial NLRP3 interaction—is important for stimulus-dependent NLRP3 activation, by bridging oligomerization. We propose that NLRP3 activation requires at least two steps (Fig. 6b). First, NLRP3 needs to be bound to NEK7, a process that is enhanced by priming and NLRP3 triggers^{13,14}. Second, the formation of the NLRP3–NEK7 complex alone may not be sufficient to induce NLRP3 activation, because NLRP3 oligomer-

ization also requires the conversion of NACHT from an inactive to an active conformation. This conformational transition may require ATP binding and other unknown allosteric triggers, as NLRP3 has previously been shown to possess ATPase activity³⁴. In the NACHT-domain-containing protein APAF-1, binding of ATP or an ATP analogue—but not ATP hydrolysis—is required for APAF-1 activation³⁵. The necessities of both NEK7 binding and the NACHT conformational change for NLRP3 activation are supported by the NEK7 requirement in mouse macrophages that contain the CAPS-associated activating mutant NLRP3(R258W) (equivalent to human NLRP3(R260W)) (Fig. 6a, Extended Data Fig. 9e) for the activation of caspase-1¹³. NLRP3 activation is complicated by many factors, including post-translational modifications^{36–39}, but our studies may have unravelled the mechanism for the NEK7-licensed NLRP3 inflammasome activation.

Our *in vitro*-reconstituted NLRP3–NEK7 complex is clearly in an inactive state and we did not observe oligomers of the purified NLRP3–NEK7 heterodimer. ATP or ATP analogues, which can induce conformational changes and oligomerization of APAF-1, did not induce oligomerization of the NLRP3–NEK7 heterodimer *in vitro*. We reasoned that this may be because NEK7 binding licenses NLRP3 but is insufficient for inflammasome assembly. Despite extensive studies, the direct triggers for the NLRP3 conformational change required to mediate the inflammasome assembly are unknown. In this regard, it is unclear whether one activated NLRP3–NEK7 complex will trigger a nucleated oligomerization, similar to the PrGJ–NAIP2–NLRC4 inflammasome^{6,19,20}, or whether the direct activator has to induce the conformational change of each NLRP3 molecule for inflammasome assembly. In plants, NLR-mediated effector-triggered immunity is often indirect, in that multiple microbial effectors may first modify a common NLR-associated protein or a specific domain in an NLR, which then relays the invasion signal to the NLR for effector-triggered immunity⁴⁰. For NLRP3, it remains to be addressed whether NEK7 can be considered a common sensor of this type.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-019-1295-z>.

Received: 21 January 2019; Accepted: 16 May 2019;

Published online 12 June 2019.

- Lamkanfi, M. & Dixit, V. M. Mechanisms and functions of inflammasomes. *Cell* **157**, 1013–1022 (2014).
- Broz, P. & Dixit, V. M. Inflammasomes: mechanism of assembly, regulation and signalling. *Nat. Rev. Immunol.* **16**, 407–420 (2016).
- Guo, H., Callaway, J. B. & Ting, J. P. Inflammasomes: mechanism of action, role in disease, and therapeutics. *Nat. Med.* **21**, 677–687 (2015).
- Shi, J., Gao, W. & Shao, F. Pyroptosis: gasdermin-mediated programmed necrotic cell death. *Trends Biochem. Sci.* **42**, 245–254 (2017).
- Liu, X. et al. Inflammasome-activated gasdermin D causes pyroptosis by forming membrane pores. *Nature* **535**, 153–158 (2016).
- Hu, Z. et al. Crystal structure of NLRC4 reveals its autoinhibition mechanism. *Science* **341**, 172–175 (2013).
- Lu, A. et al. Unified polymerization mechanism for the assembly of ASC-dependent inflammasomes. *Cell* **156**, 1193–1206 (2014).
- Lu, A. et al. Molecular basis of caspase-1 polymerization and its inhibition by a new capping mechanism. *Nat. Struct. Mol. Biol.* **23**, 416–425 (2016).
- Lu, A. et al. Plasticity in PYD assembly revealed by cryo-EM structure of the PYD filament of AIM2. *Cell Discov.* **1**, 15013 (2015).
- Rathinam, V. A., Vanaja, S. K. & Fitzgerald, K. A. Regulation of inflammasome signaling. *Nat. Immunol.* **13**, 333–342 (2012).
- Muñoz-Planillo, R. et al. K⁺ efflux is the common trigger of NLRP3 inflammasome activation by bacterial toxins and particulate matter. *Immunity* **38**, 1142–1153 (2013).

- Chen, J. & Chen, Z. J. PtdIns4P on dispersed *trans*-Golgi network mediates NLRP3 inflammasome activation. *Nature* **564**, 71–76 (2018).
- He, Y., Zeng, M. Y., Yang, D., Motro, B. & Núñez, G. NEK7 is an essential mediator of NLRP3 activation downstream of potassium efflux. *Nature* **530**, 354–357 (2016).
- Shi, H. et al. NLRP3 activation and mitosis are mutually exclusive events coordinated by NEK7, a new inflammasome component. *Nat. Immunol.* **17**, 250–258 (2016).
- Schmid-Burgk, J. L. et al. A genome-wide CRISPR (clustered regularly interspaced short palindromic repeats) screen identifies NEK7 as an essential component of NLRP3 inflammasome activation. *J. Biol. Chem.* **291**, 103–109 (2016).
- Shi, H., Murray, A. & Beutler, B. Reconstruction of the mouse inflammasome system in HEK293T cells. *Bio Protoc.* **6**, e1986 (2016).
- Coll, R. C. et al. A small-molecule inhibitor of the NLRP3 inflammasome for the treatment of inflammatory diseases. *Nat. Med.* **21**, 248–255 (2015).
- Dar, A. C., Dever, T. E. & Sicheri, F. Higher-order substrate recognition of eIF2 α by the RNA-dependent protein kinase PKR. *Cell* **122**, 887–900 (2005).
- Hu, Z. et al. Structural and biochemical basis for induced self-propagation of NLRC4. *Science* **350**, 399–404 (2015).
- Zhang, L. et al. Cryo-EM structure of the activated NAIP2–NLRC4 inflammasome reveals nucleated polymerization. *Science* **350**, 404–409 (2015).
- Maekawa, S., Ohto, U., Shibata, T., Miyake, K. & Shimizu, T. Crystal structure of NOD2 and its implications in human disease. *Nat. Commun.* **7**, 11813 (2016).
- Yang, X. et al. Structural basis for specific flagellin recognition by the NLR protein NAIP5. *Cell Res.* **28**, 35–47 (2018).
- Tenthorey, J. L. et al. The structural basis of flagellin detection by NAIP5: a strategy to limit pathogen immune evasion. *Science* **358**, 888–893 (2017).
- Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N. & Sternberg, M. J. The Pyre2 web portal for protein modeling, prediction and analysis. *Nat. Protoc.* **10**, 845–858 (2015).
- Richards, M. W. et al. An autoinhibitory tyrosine motif in the cell-cycle-regulated Nek7 kinase is released through binding of Nek9. *Mol. Cell* **36**, 560–570 (2009).
- Qu, Y. et al. Phosphorylation of NLRC4 is critical for inflammasome activation. *Nature* **490**, 539–542 (2012).
- Suzuki, S. et al. *Shigella* type III secretion protein Mxi1 is recognized by Naip2 to induce Nlr4 inflammasome activation independently of Pkc δ . *PLoS Pathog.* **10**, e1003926 (2014).
- Krissinel, E. & Henrick, K. Inference of macromolecular assemblies from crystalline state. *J. Mol. Biol.* **372**, 774–797 (2007).
- Haq, T. et al. Mechanistic basis of Nek7 activation through Nek9 binding and induced dimerization. *Nat. Commun.* **6**, 8771 (2015).
- Spalinger, M. R. et al. NLRP3 tyrosine phosphorylation is controlled by protein tyrosine phosphatase PTPN22. *J. Clin. Invest.* **126**, 1783–1800 (2016).
- Hafner-Bratkovič, I. et al. NLRP3 lacking the leucine-rich repeat domain can be fully activated via the canonical inflammasome pathway. *Nat. Commun.* **9**, 5182 (2018).
- Moghaddas, F. et al. Autoinflammatory mutation in NLRC4 reveals a leucine-rich repeat (LRR)–LRR oligomerization interface. *J. Allergy Clin. Immunol.* **142**, 1956–1967.e6 (2018).
- Touitou, I. et al. Infervers: an evolving mutation database for auto-inflammatory syndromes. *Hum. Mutat.* **24**, 194–198 (2004).
- Duncan, J. A. et al. Cryopyrin/NALP3 binds ATP/dATP, is an ATPase, and requires ATP binding to mediate inflammatory signaling. *Proc. Natl Acad. Sci. USA* **104**, 8041–8046 (2007).
- Jiang, X. & Wang, X. Cytochrome c promotes caspase-9 activation by inducing nucleotide binding to Apaf-1. *J. Biol. Chem.* **275**, 31199–31203 (2000).
- Song, N. & Li, T. Regulation of NLRP3 inflammasome by phosphorylation. *Front. Immunol.* **9**, 2305 (2018).
- Barry, R. et al. SUMO-mediated regulation of NLRP3 modulates inflammasome activity. *Nat. Commun.* **9**, 3001 (2018).
- Py, B. F., Kim, M. S., Vakifahmetoglu-Norberg, H. & Yuan, J. Deubiquitination of NLRP3 by BRCC3 critically regulates inflammasome activity. *Mol. Cell* **49**, 331–338 (2013).
- Mangan, M. S. J. et al. Targeting the NLRP3 inflammasome in inflammatory diseases. *Nat. Rev. Drug Discov.* **17**, 688 (2018).
- Khan, M., Subramaniam, R. & Desveaux, D. Of guards, decoys, baits and traps: pathogen perception in plants by type III effector sensors. *Curr. Opin. Microbiol.* **29**, 49–55 (2016).
- Kucukelbir, A., Sigworth, F. J. & Tagare, H. D. Quantifying the local resolution of cryo-EM density maps. *Nat. Methods* **11**, 63–65 (2014).

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

METHODS

Cloning, expression, purification and reconstitution of recombinant human NLRP3 and NEK7 complex. Full-length (amino acids 1–302) and N-terminally truncated (amino acids 31–302) human NEK7 sequences were cloned into a modified pET28a vector with an N-terminal 6×His-SUMO tag followed by the Ulp1 protease site. These constructs were transfected into *Escherichia coli* BL21 (DE3) cells, which were grown in LB medium supplemented with 50 µg/ml kanamycin. Protein expression was induced overnight at 18 °C with 0.2 mM isopropyl-β-D-thiogalactopyranoside after optical density at 600 nm reached 0.8. For N-terminally truncated NEK7, cells were collected and resuspended in a lysis buffer containing 20 mM Tris-HCl at pH 7.4, 300 mM NaCl, 0.5 mM Tris(2-carboxyethyl) phosphine hydrochloride (TCEP) and 20 mM imidazole. For the full-length NEK7, the lysis was performed in buffer containing 50 mM Tris-HCl at pH 7.5, 500 mM NaCl, 5 mM MgCl₂, 10 mM imidazole, 10% glycerol and 2 mM β-mercaptoethanol. The proteins were first purified by affinity chromatography using Ni-NTA beads (Qiagen). The 6×His-SUMO tag was cleaved by overnight Ulp1 protease digestion with dialysis at 4 °C and removed by passing through Ni-NTA beads again. The NEK7 proteins with SUMO-tag or after protease cleavage were further purified by size-exclusion chromatography on a Superdex 200 column (GE Healthcare Life Sciences) equilibrated with the gel filtration buffer containing 20 mM Tris-HCl at pH 7.5, 150 mM NaCl and 0.5 mM TCEP.

An artificial NEK7 dimer was engineered based on the protein kinase R dimer structure (PDB code 2A19)¹⁸ by a back-to-back interaction at the N-terminal lobe. The N-terminal disordered region (residues 1–33) of NEK7 was replaced by the non-canonical helix α0 (residues 258–266) of protein kinase R, and several surface residues at the predicted dimerization interface were mutated to corresponding residues in protein kinase R (L54R, V58K, P59T, K87A, A99V, S100C, E103T and D104G). Protein expression and purification of the NEK7 dimer followed the same procedure as full-length or N-terminally truncated wild-type NEK7.

NLRP3 with the pyrin domain deleted (amino acids 134–1034) was cloned into a modified pFastBac vector with an N-terminal MBP tag. The sequence 134-YRKKYRK-140 was changed to 134-YCAKYRA-140, designed to avoid a potential clash with the N-terminal MBP. The baculovirus of NLRP3 was prepared using the Bac-to-Bac system (Invitrogen). To express the protein, 10 ml of the NLRP3 baculovirus was used to infect 1 l Sf9 cells, which were collected 48 h after infection. Cells were lysed by sonication in buffer containing 20 mM Tris-HCl at pH 7.5, 200 mM NaCl, 0.5 mM TCEP and 10% glycerol with freshly added protease inhibitor cocktail (Sigma). After centrifugation, the supernatant was incubated with 3 ml amylose resin at 4 °C for 1 h. The mixture was then subjected to gravity flow and the bound MBP-tagged NLRP3 protein was eluted with 25 mM maltose. The eluted protein was further purified by size-exclusion chromatography on a Superose 6 column (GE Healthcare Life Sciences) in the same gel filtration buffer used for NEK7.

The complexes of MBP-tagged NLRP3 with NEK7 monomer and dimer were reconstituted by size-exclusion chromatography on a Superose 6 column equilibrated with the gel filtration buffer. The eluted proteins were analysed on SDS-PAGE gels for homogeneity and stoichiometry. The fractions were then pooled and concentrated for further analysis and cryo-EM data collection.

All mutagenesis experiments were performed by the QuikChange site-directed mutagenesis protocol using Q5 High-Fidelity 2× Master Mix (NEB). All plasmids were verified by DNA sequencing.

Microscale thermophoresis. Proteins were labelled with the cysteine-reactive dye Alexa Fluor 488 C5 maleimide (Thermo Fisher). Purified NLRP3 was incubated with ~20-fold molar excess of the dye in gel-filtration buffer for 1 h. After labelling, the excess dye was removed by applying the sample on Superdex 200 column equilibrated with the gel filtration buffer. Tween 20 (0.05%) was added to the protein buffer for microscale thermophoresis measurements before the experiment.

Datasets were collected at a temperature of 25.1 °C. For the Alexa Fluor 488-labelled NLRP3 interaction with NEK7, a concentration series of NEK7 was prepared using a 1:1 serial dilution of NEK7 in gel filtration buffer supplemented with 0.05% Tween 20. The range of NEK7 concentration used was from 32.5 µM to a final concentration of 1 nM, over 16 serial diluted capillaries with 10-µl samples. The interaction was initiated by the addition 10 µl of 140 nM Alexa Fluor 488-labelled NLRP3 to each reaction mixture, resulting in a 70 nM final concentration of NLRP3. The LED power was 100% and the microscale thermophoresis power (that is, the power supplied to the infrared laser) was 60%. The pre-microscale thermophoresis period was 5 s, the microscale thermophoresis acquisition period was 20 s, and the post-microscale thermophoresis period was 5 s. Data were analysed by MO Control software provided by NanoTemper. For analysis, the 32.5 µM concentration measurement was deleted as an outlier.

Multi-angle light scattering. For molecular-mass determination by multi-angle light scattering (MALS), protein samples were injected into a Superdex 200 (10/300 GL) gel-filtration column equilibrated with the gel filtration buffer. The chromatography system was attached to a three-angle light scattering detector (miniDAWN

TRISTAR) and a refractive index detector (Optilab DSP) (Wyatt Technology). Data were collected every 0.5 s with a flow rate of 0.5 ml/min. Data analysis was carried out using ASTRA V.

Negative-staining electron microscopy. For negative staining, 4 µl of an NLRP3-NEK7 complex was placed onto a glow-discharged copper grid (Electron Microscopy Sciences) coated with a layer of thin carbon, washed twice with H₂O, stained with 1–2% uranyl formate for 30 s and air-dried. The grids were imaged on a Tecnai G2 Spirit BioTWIN electron microscope and recorded with an AMT 2k CCD camera (Harvard Medical School Electron Microscopy Facility).

Cryo-EM data collection. An NLRP3-NEK7 complex at 0.7 mg/ml was incubated with 0.3 mM ADP, 5 mM MgCl₂ and 0.3 mM MCC950 on ice for 30 min. A 3-µl drop of the NLRP3-NEK7 complex was applied to a glow-discharged Quantifoil grid (R 1.2/1.3 400 mesh, copper, Electron Microscopy Sciences), blotted for 3–3.5 s in 100% humidity at 4 °C and plunged into liquid ethane using an FEI Vitrobot Mark IV. For data collection on an FEI Talos Arctica microscope operating at an acceleration voltage of 200 keV, movies were acquired on a K2 Summit direct electron detector (Gatan) in super-resolution counting mode, with 10.03 s total exposure time over 59 frames and an accumulated dose of 58.1 electrons per Å². The super-resolution pixel size was 0.5843 Å. The defocus level in the data collection was set in the range of –1.0 to –3.0 µm. A total of 3,667 movies were collected.

For data collection on an FEI Titan Krios G2 microscope operating at 300 keV, the cryo-grids were visually inspected for ice contaminations and flatness before loading into the microscope autoloader. Data collection used post-column BioQuantum energy filter (Gatan) in zero-loss imaging mode with a 20-µm energy slit and K2 Summit direct electron detector (Gatan) in super-resolution counting mode (Peking University Electron Microscopy Laboratory and Cryo-EM Platform). Although the grids appeared to have icy areas, there were plenty of good holes, which yielded images that were free of strong diffraction patterns corresponding to crystalline ice. Coma-free alignment was manually optimized before data collection; the parallel-beam illumination area under a nanoprobe mode was selected to achieve the optimal dose rate, avoid exposure interference and minimize edge fringes on the image. To increase collection efficiency, the data were collected in part with multiple exposures per grid hole, during which the image shift applied was no more than 250 nm. The data collection process was set up and carried out in SerialEM⁴². Movies were acquired with a dose rate of 5.4 electrons per second per physical pixel, and a total dose of 54.8 electrons per Å² equally distributed in 7.2 s to 40 frames. The super-resolution pixel size was 0.42 Å. A total of 15,861 movies were collected during 5 separate sessions, among which 1,340 movies were collected with a moderate stage tilt of 20°. The defocus levels of the images were set in the range of –0.8 to –3.0 µm.

Cryo-EM data processing. For Arctica data processing, raw movies were corrected by gain reference and for beam-induced motion, and summed into motion-corrected images using MotionCor2⁴³. The CTFFIND4 program⁴⁴ was used to determine the actual defocus level of each micrograph. RELION 3.0⁴⁵ and cisTEM⁴⁶ were used for subsequent image processing (Extended Data Fig. 2). First, template-free autopicking was used to generate an initial particle set for 2D classification. Good 2D classes were then used as templates for additional autopicking. Almost one million particles were picked which were subjected to multiple rounds of 2D classification until the classes looked homogenous. A few good class averages in different orientations were selected for the reconstruction of an initial model, which was low-pass-filtered to 60.0 Å to use as the input reference for 3D classification of 522,320 particles. One selected 3D class was further classified and refined to reach an overall resolution of 4.3 Å, measured by gold-standard Fourier shell correlation (FSC) between half maps (Extended Data Table 1). Owing to particle orientation preference, the map suffers from anisotropic resolution.

Krios data processing also used MotionCor2⁴³ for aligning and summing movie frames at the pixel size 0.84 Å. The averaged images were binned threefold to obtain TIFF images for visual screening. Bad images, such as those with excessive ice contaminants, were removed. The contrast-transfer function parameters of each drift-corrected micrograph were then calculated using the program Gctf⁴⁷. Particles were automatically picked from twofold-binned micrographs using Gautomatch (<https://sbgrid.org/software/titles/gautomatch>) and a Gaussian sphere template, which yielded a total of 1,850,332 particles. Reference-free 2D and 3D classifications were carried out with both RELION 3.0⁴⁵ and ROME 1.1.2⁴⁸. The latter combines maximum-likelihood-based image alignment with statistical machine-learning-based classification⁴⁸, and can output a large number of refined 2D classes with subtle differences. The NLRP3-NEK7 complex particles naturally and preferentially orientated with the curved LRR plane on the grid plane. To alleviate this orientation preference, we first included a subset of data that was collected with a modest 20° stage tilt. Because of the large vertical drift of carbon grids projected in the lateral direction, we could not use a larger tilt. We also retained as many particles as possible from the rare views that may not have been previously included in RELION-based 2D classification. This was done by multiple

rounds of deep unsupervised 2D classifications with the ROME 1.1.2 package⁴⁸, which allowed us to improve the conformational homogeneity of selected particles. Three-dimensional classification and refinement were carried out in Relion 3.0⁴⁵. Map reconstruction was done with ROME 1.1.2⁴⁸, and the local resolution map was produced with ResMap⁴¹, which is embedded in RELION 3.0⁴⁵ (Extended Data Fig. 3). The inclusion of the tilted data and deep 2D classification spread out the orientation distribution and helped to improve the overall features, such as the connecting loop density on both sides of the LRR region. Data processing statistics are summarized in Extended Data Table 1.

Model building and display. NLRP3 initial homology model was generated using the Phyre2 server²⁴. Manual adjustment, rigid-body fitting, flexible fitting and segment-based real-space refinement were performed in distinct parts of the initial model to fit in the density in Coot⁴⁹, with help of Chimera⁵⁰ and real-space refinement in Phenix⁵¹. In the fitting to the final 3.8 Å map, large side-chain densities were used to guide the local register of the backbone trace. For example, W774 and W831 flanking the inner surface of the LRR domain were used to confirm the residue register in LRR. The density that we observed accounts for NLRP3 residues (amino acids 133–1034) that include the NBD, HD1, WHD, HD2 and LRR domains. A few unstructured regions, including part of the HD2 domain (amino acids 654–683), were omitted owing to poor density. The model was fitted and modelled to the best of our knowledge. The remaining part of the cryo-EM map was fitted with the NEK7 crystal structure (PDB code 2WQN)²⁵. Only NEK7 C-terminal lobe residues (amino acids 113–300) were fitted in the density because the N-lobe density appeared to have been averaged out in the data processing, probably owing to flexibility. Interaction analysis was conducted using PISA²⁸ and structure representations were generated in Chimera⁵⁰, Pymol⁵² and ResMap⁴¹.

Thermal shift assay. Purified 2 µM NLRP3 or NEK7 was mixed with 1-fold protein thermal shift dye (Thermo fisher scientific) and 200 µM MCC950 or ADP, or 100 µM MCC950 plus 100 µM ADP in a 20-µl reaction volume. Thermal scanning (25 to 75 °C at 1.5 °C per minute) was performed, and melting curves were recorded on a StepOne RT-PCR machine. Data analysis was done by Protein Thermal Shift software (Thermo Fisher Scientific).

In vitro pull-downs. MBP-tagged NLRP3 (1.5 µM) was mixed with 3 µM full-length wild-type or mutant NEK7 in buffer containing 30 mM HEPES at pH 7.5, 150 mM NaCl and 0.5 mM TCEP, and incubated for 1 h on ice. The mixture was further incubated for 1 h with 30 µl amylose resin and washed twice with 250 µl of the same buffer, followed by 1 h elution with 50 mM maltose. Input and elution fractions were analysed with SDS-PAGE. NEK7 and NLRP3 bands were quantified with ImageJ and the efficiency of binding was calculated from the ratio of NEK7 to NLRP3 band intensities. Three independent experiments were performed, and the efficiency of binding was presented as mean ± standard deviation.

Generation of stable cell lines and cellular reconstitution. To reconstitute NLRP3-knockout iBMDMs with wild-type or mutant full-length mNG-tagged human NLRP3, lentiviral particles were generated as follows. On day 0, the lentivirus was produced using HEK293T (60% confluent) cells by co-transfecting 1 µg of a pLV vector containing the wild-type or mutant NLRP3, 750 ng psPAX2 packaging plasmid and 250 ng pMD2.G envelope plasmid. The packaging and envelope plasmids were gifts from D. Trono (Addgene numbers 12260 and 12259, respectively). The transfected cells were incubated overnight. On day 1, the medium was removed, and replenished with 1 ml fresh medium. The cells were incubated for another day. The expression was analysed by a confocal laser scanning microscope FluoView FV1000 (Olympus America), equipped with 488 argon for green excitation.

On day 2, the supernatant containing the virus was filtered using a 0.45-µm filter and used directly to infect NLRP3-knockout iBMDMs (2.5×10^6 cells/ml) using a spinfection protocol to increase the efficacy. Spinfection was performed at 2,500g for 90 min at 37 °C using 8 µg/ml polybrene (Santa Cruz Biotechnology, cat. no. sc-134220). After spinfection, cells were further incubated for the expression of reconstituted proteins for 48 h. Reconstitution was verified by western blotting using a specific anti-human NLRP3 antibody (Adipogen, cat. no. AG-20B-0014-C100).

Inflammasome activation assays. NLRP3-knockout iBMDMs reconstituted with human wild-type or mutant NLRP3 were seeded at a density of 3×10^6 cells/ml. On the next day, the cells were primed with 1 µg/ml LPS (Invivogen, cat. no. trl-b5lps) for 4 h. Post priming, the cells were activated with 20 µM nigericin (Sigma-Aldrich, cat. no. N7143-5MG) for 30 min, and the supernatant was collected. Released IL-1β was analysed using an ELISA kit (Affymetrix eBioscience, cat. no. 88-7013), and released LDH was measured using the LDH Glo cytotoxicity assay (Promega, cat. no. J2380), according to manufacturer's instructions. For analysis of caspase-1 cleavage, the whole-cell lysate was prepared in 1 × laemmli sample buffer, resolved on 12.5% SDS-PAGE and western-blotted using a specific anti-mouse caspase-1 antibody (Adipogen, cat. no. AG-20B-0042-C100). The graphs were plotted using GraphPad Prism 7 software.

Immunolabelling and antibodies. Reconstituted iBMDM cells were primed for 4 h with LPS, and then activated with nigericin for 30 min. Fixative, permeabi-

lization and blocking buffers were prepared in Brinkley Buffer 80 (BRB80), and kept at 37 °C before use. BRB80 buffer was prepared freshly using 80 mM PIPES, 1 mM MgCl₂, 1 mM EGTA, titrated to pH 6.8 with a saturated solution of KOH.

Cells were fixed in 3.7% paraformaldehyde for 5 min at room temperature. Afterwards, cells were washed twice using BRB80 with 5-min intervals between washes. Permeabilization was carried out for 5 min at room temperature using 0.15% Triton X-100 (in 1 × BRB80). Washing was carried out to remove permeabilization buffer. Cells were then blocked for 1 h at room temperature using blocking buffer (3% gelatin from cold water fish skin prepared in 1 × BRB80). Cells were incubated with primary antibody for overnight (anti-ASC, 1:1,000, Cell Signaling Technology, cat. no. 67824S). The cells were washed and incubated with secondary antibody for an hour at room temperature (Goat anti-rabbit (H+L), Alexa Fluor 647 HRP conjugated 1:1,000 Thermo Fisher Scientific, cat. no. A-21245). The nucleus was stained using Hoechst 33342 (Invitrogen, cat. no. H3750). In between each step, extensive washing steps were carried out to remove unbound antibodies and stains.

Confocal laser-scanning microscopy. After activation, reconstituted iBMDMs were washed once with 1 × BRB buffer, fixed and labelled with antibodies as mentioned in 'Immunolabelling and antibodies'. Confocal sections were obtained with an Olympus confocal laser scanning microscope FluoView FV1000 (Olympus America), equipped with 405 diode (for blue excitation) or a 635 diode (for far-red excitation). Images were captured using 40 × (0.95 NA, air objective), with Olympus FluoView version 3.0 viewer software. The images were acquired identically, and processed using Adobe Photoshop software.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

The cryo-EM map has been deposited in the Electron Microscopy Data Bank under the accession number EMD-0476. The atomic coordinates have been deposited in the Protein Data Bank under the accession number 6NPY. All other data can be obtained from the corresponding author upon reasonable request.

42. Mastronarde, D. N. Automated electron microscope tomography using robust prediction of specimen movements. *J. Struct. Biol.* **152**, 36–51 (2005).
43. Zheng, S. Q. et al. MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy. *Nat. Methods* **14**, 331–332 (2017).
44. Rohou, A. & Grigorieff, N. CTFFIND4: Fast and accurate defocus estimation from electron micrographs. *J. Struct. Biol.* **192**, 216–221 (2015).
45. Zivanov, J. et al. New tools for automated high-resolution cryo-EM structure determination in RELION-3. *eLife* **7**, e42166 (2018).
46. Grant, T., Rohou, A. & Grigorieff, N. cisTEM, user-friendly software for single-particle image processing. *eLife* **7**, e35383 (2018).
47. Zhang, K. Gctf: real-time CTF determination and correction. *J. Struct. Biol.* **193**, 1–12 (2016).
48. Wu, J. et al. Massively parallel unsupervised single-particle cryo-EM data clustering via statistical manifold learning. *PLoS ONE* **12**, e0182130 (2017).
49. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D* **66**, 486–501 (2010).
50. Pettersen, E. F. et al. UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).
51. Adams, P. D. et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66**, 213–221 (2010).
52. Delano, W. L. The PyMol Molecular Graphics System <https://pymol.org/2/> (2002).
53. Scheres, S. H. RELION: implementation of a Bayesian approach to cryo-EM structure determination. *J. Struct. Biol.* **180**, 519–530 (2012).
54. Robert, X. & Gouet, P. Deciphering key features in protein structures with the new ENDScript server. *Nucleic Acids Res.* **42**, W320–W320 (2014).

Acknowledgements We thank the University of Massachusetts Cryo-EM Core Facility and the National Cryo-EM Facility at the National Cancer Institute for 200-keV data collection, the Electron Microscopy Laboratory and Cryo-EM Platform at Peking University for 300-keV data collection, and C. Xu, K. Song, K. Lee, X. Li, Y. Ma and D. Yu for technical support. This work was funded in part by the US NIH grant DP1HD087988 (H.W.), R01AI124491 (H.W.), R01AI06331 (G.N.), an Intel Corporation academic grant (Y.M.), the Thousand Talents Plan of China (Y.M.), National Natural Science Foundation of China grant 11774012 (Y.M.) and Natural Science Foundation of Beijing City grant Z180016 (Y.M.). Data processing was performed in part in the Sullivan cluster, which is funded in part by a gift from Mr. and Mrs. D. J. Sullivan Jr, and in the Weiming No.1 and Life Science No. 1 High-Performance Computing Platform at Peking University. We thank H. Leung (PCMM, BCH confocal core facility manager).

Author contributions H.W., L.W. and H.S. conceived the study. L.W. designed constructs. L.W., Q.Q. and H.S. purified the complexes, and H.S., L.W. and Q.Q. made cryo-grids for data collection. H.S., L.W., W.L.W., Q.Q., Z.W. and Y.M. collected data. H.S., W.L.W. and Q.Q. analysed cryo-EM data, and H.W. and Y.M. supervised data processing. H.S. performed initial model

building and refinement, and W.L.W. and Y.M. performed additional model fitting and refinement. H.S. designed mutants for in vitro and cell-based assays. V.G.M. performed all cell-based assays. L.A. and A.V.H. performed NEK7-mutant assays. G.N. provided reagents for cell-based assays and valuable discussions. H.W., H.S. and Y.M. wrote the manuscript, and all authors provided comments on the manuscript.

Competing interests L.W. and A.V.H. are employees, and H.W. is co-founder, of SMOC Therapeutics; the other authors declare no competing interests.

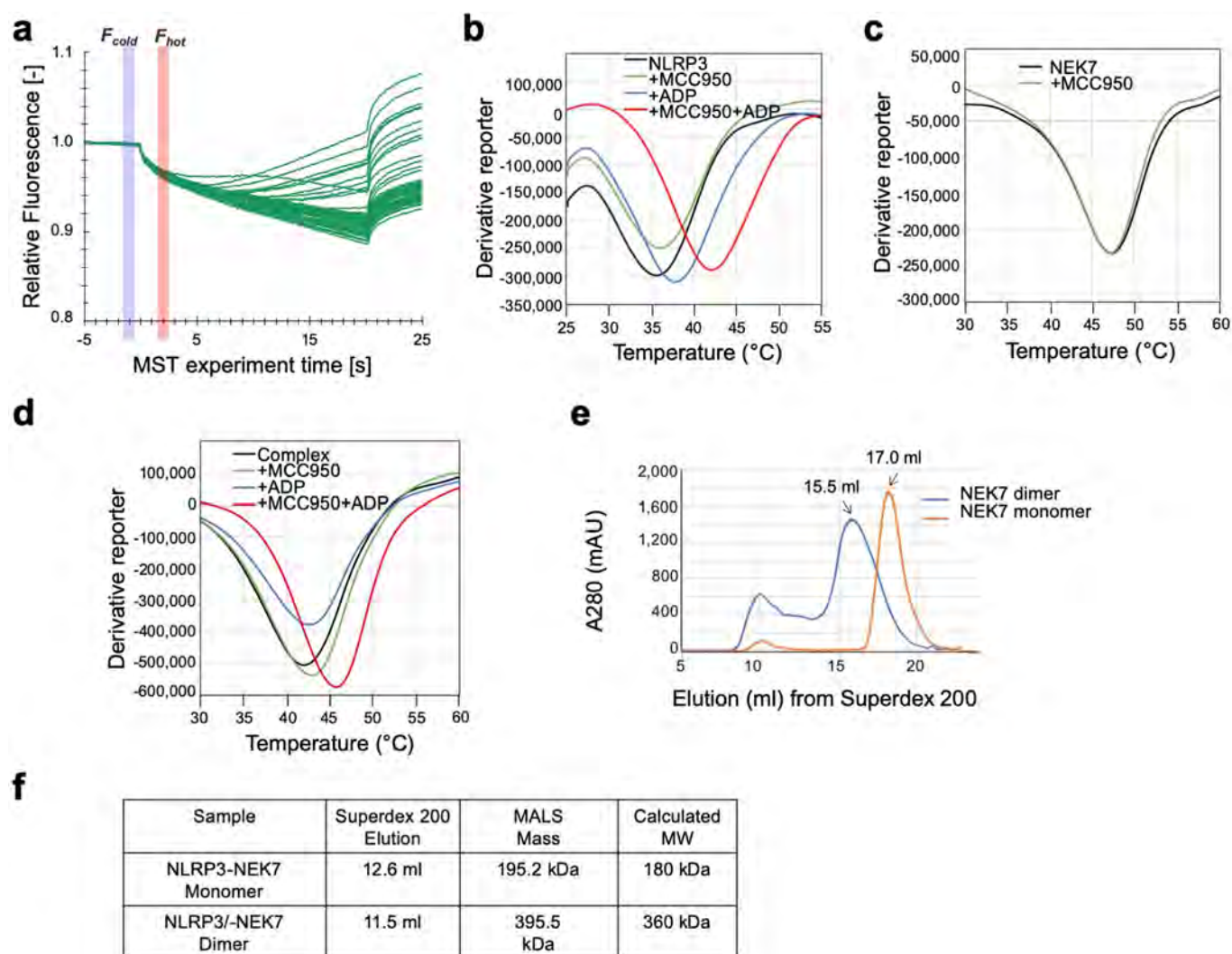
Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-019-1295-z>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

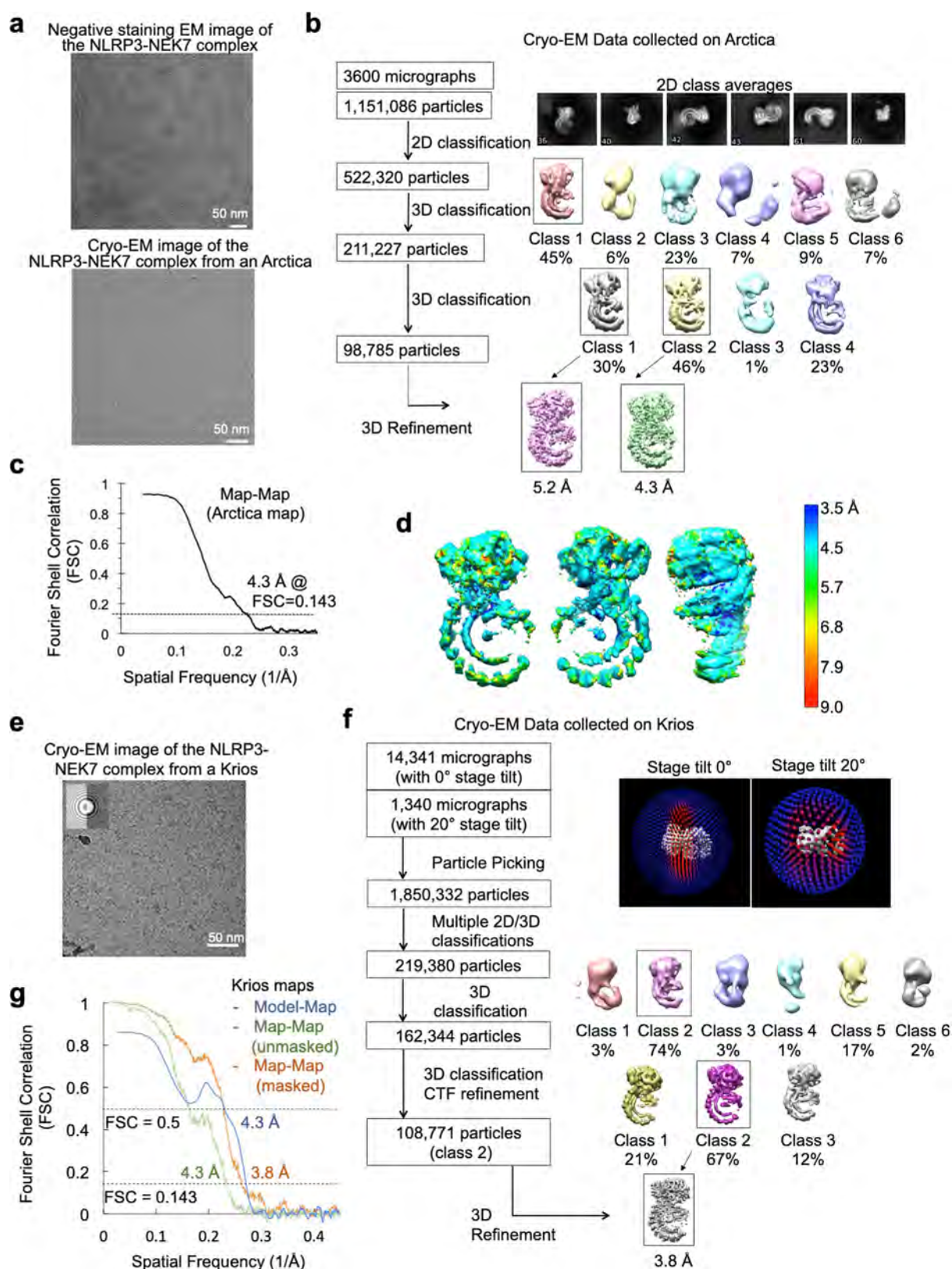
Correspondence and requests for materials should be addressed to Y.M. or H.W.

Peer Reviewer information: *Nature* thanks Eicke Latz, Edward Miao and the other anonymous reviewer(s) for their contribution to the peer review of this work.



Extended Data Fig. 1 | NLRP3-NEK7 protein purification and characterization. **a**, Raw traces (of three independent replicates) of microscale thermophoresis measurements corresponding to titration of NEK7 against Alexa Fluor 488-labelled NLRP3. **b-d**, Representative thermal denaturation curves of NLRP3 (**b**), NEK7 (**c**) and complex (**d**), alone and in the presence of ADP and/or MCC950. The peak minima

of the derivative curves correspond to the protein melting temperatures (T_m) (repeated ≥ 3 times). **e**, Gel filtration profile of wild-type NEK7 (monomer) and engineered NEK7 dimer on Superdex 200 column (repeated ≥ 5 times). **f**, Molecular masses of NLRP3-NEK7 dimer and monomer complexes, measured by in-line MALS.

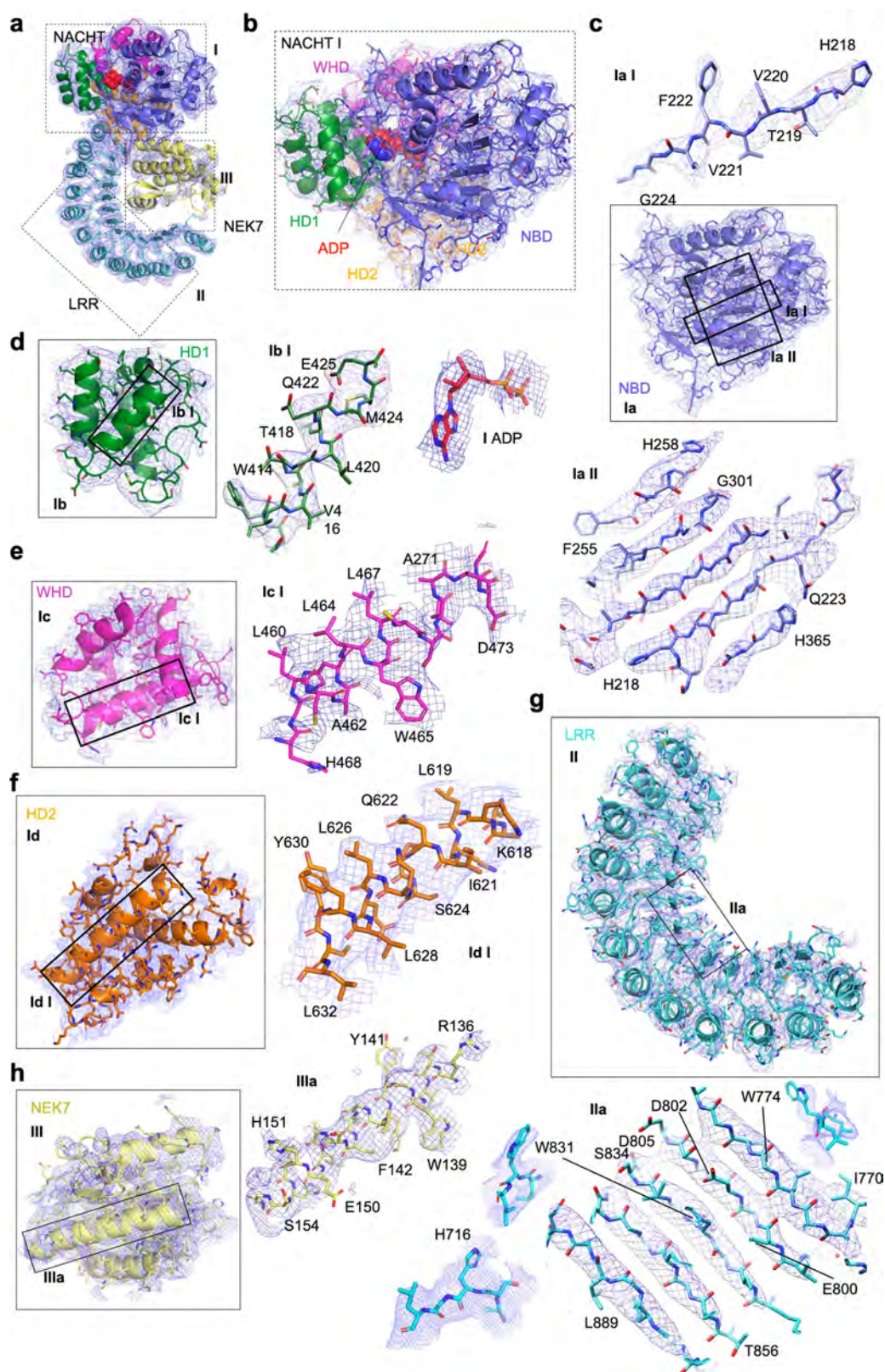


Extended Data Fig. 2 | Analysis and workflow for cryo-EM data

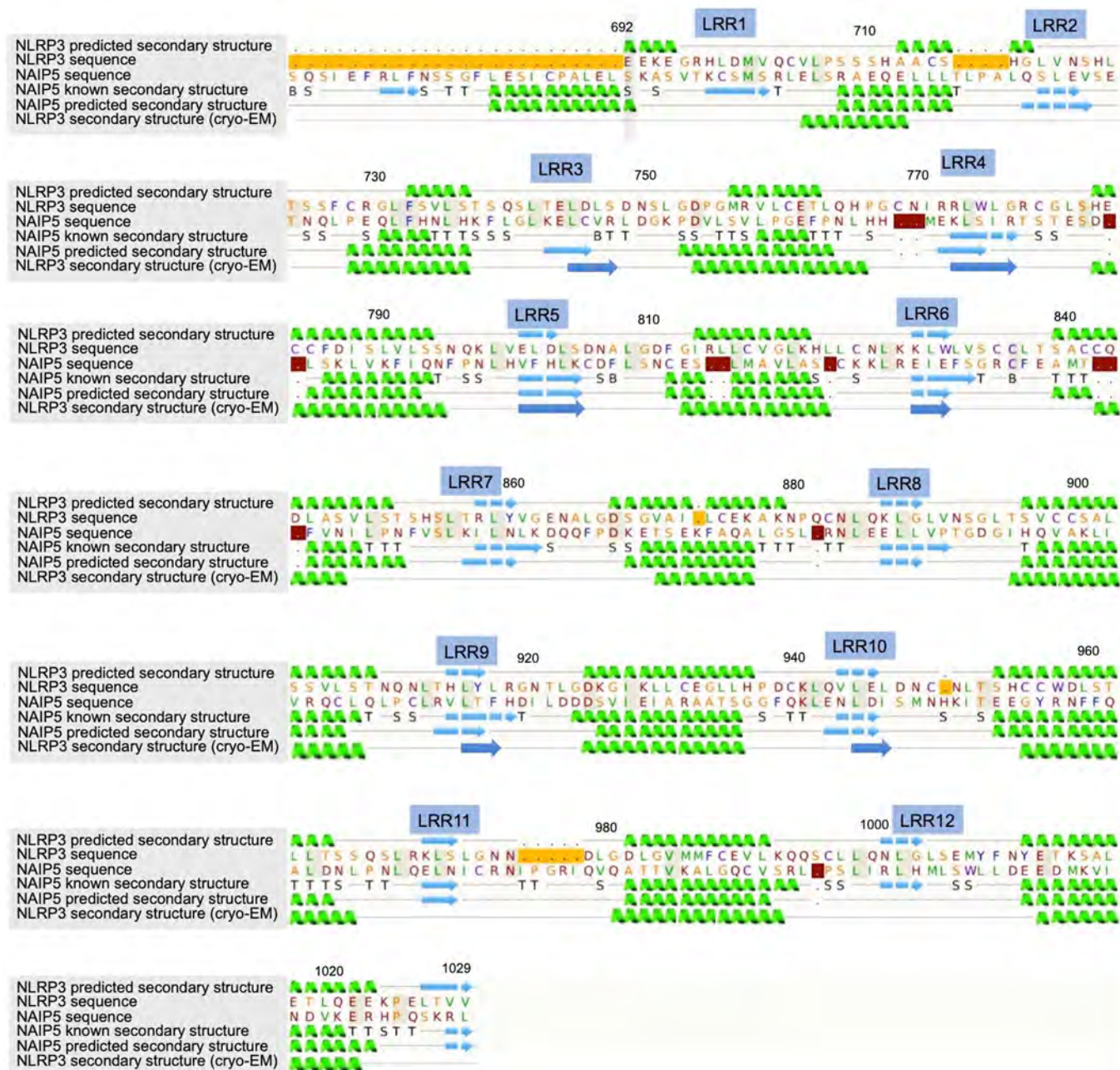
collected on a Talos Arctica and on a Titan Krios. **a**, A negative-staining electron microscopy image (top, one among a few hundred images) and a cryo-EM micrograph (bottom, one among a few thousand images) of the NLRP3-NEK7 complex from a Talos Arctica. Scale bars, 50 nm.

b, Workflow of cryo-EM data analysis of the Arctica data, performed in RELION 3.0⁵³ and cisTEM⁴⁶. **c**, Gold-standard FSC curve between two half maps from the Arctica data. **d**, Local resolution estimation of the Arctica map generated by ResMap⁴¹, coloured on the cryo-EM density

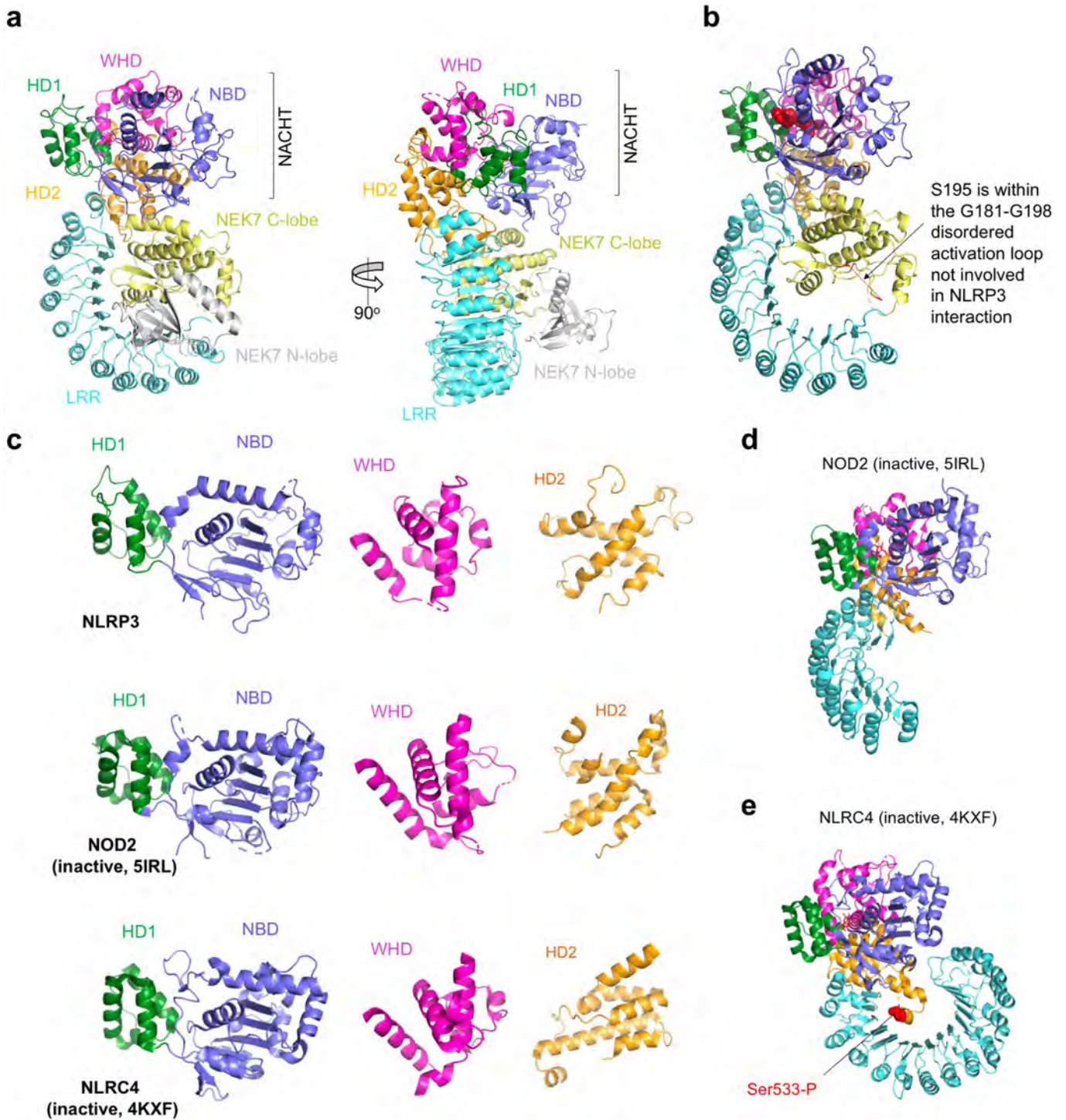
(8 σ). The highest resolution is observed where NEK7 interacts with NLRP3. **e**, A cryo-EM micrograph of the NLRP3-NEK7 complex from a Titan Krios (one among more than 10,000 images). Scale bar, 50 nm. The inset shows the modelled (left) and actual (right) Thon rings. **f**, Workflow of cryo-EM data analysis of the Krios data, done in RELION 3.0⁵³ and ROME 1.1⁴⁸. The top right insets show the orientation distributions of the particles from the dataset of tilt 0° and tilt 20°. **g**, Gold-standard FSC curves between two half maps from the Krios data with mask (orange) and without mask (green), and between map and model (blue).



Extended Data Fig. 3 | Local density fitted with NLRP3 and NEK7. All magnified views are labelled with domain names and selected segment residue numbers. Densities are shown at 3σ .

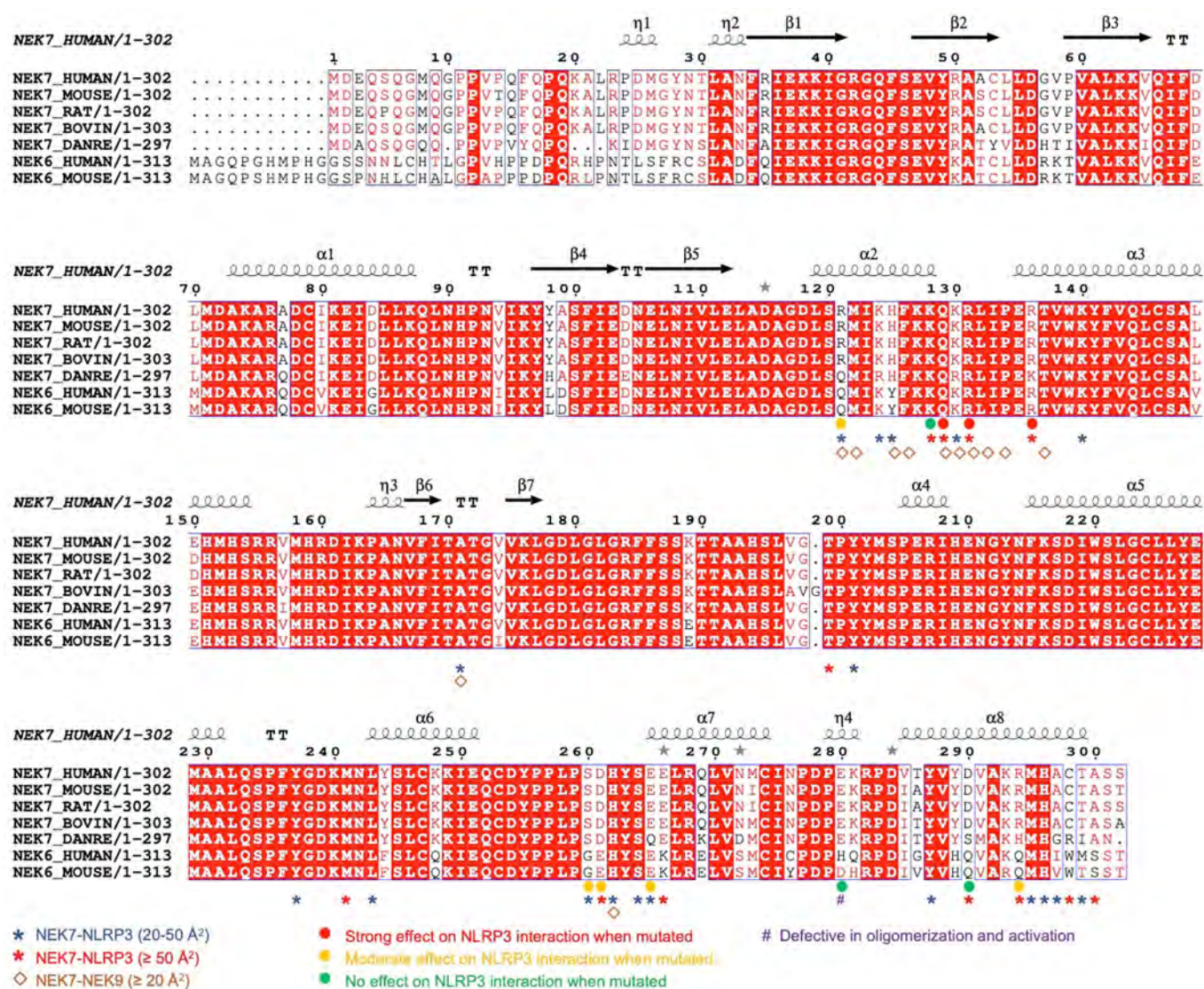


Extended Data Fig. 4 | Predicted and cryo-EM-derived secondary structures of the NLRP3-LRR domains aligned to NLRP3 and NAIP5 sequences. Twelve LRRs in NLRP3 were predicted by the Phyre2 server²⁴.



Extended Data Fig. 5 | The structure of the NLRP3-NEK7 complex, showing the modelled full-length NEK7 and NLRP3 domain comparisons. a, Two views of the NLRP3-NEK7 complex structure with the NEK7 N-lobe (grey) modelled from the NEK7 crystal structure (PDB code 2WQM). **b,** The disordered activation loop, including S195 in the NLRP3-NEK7 complex, which is not involved in NLRP3

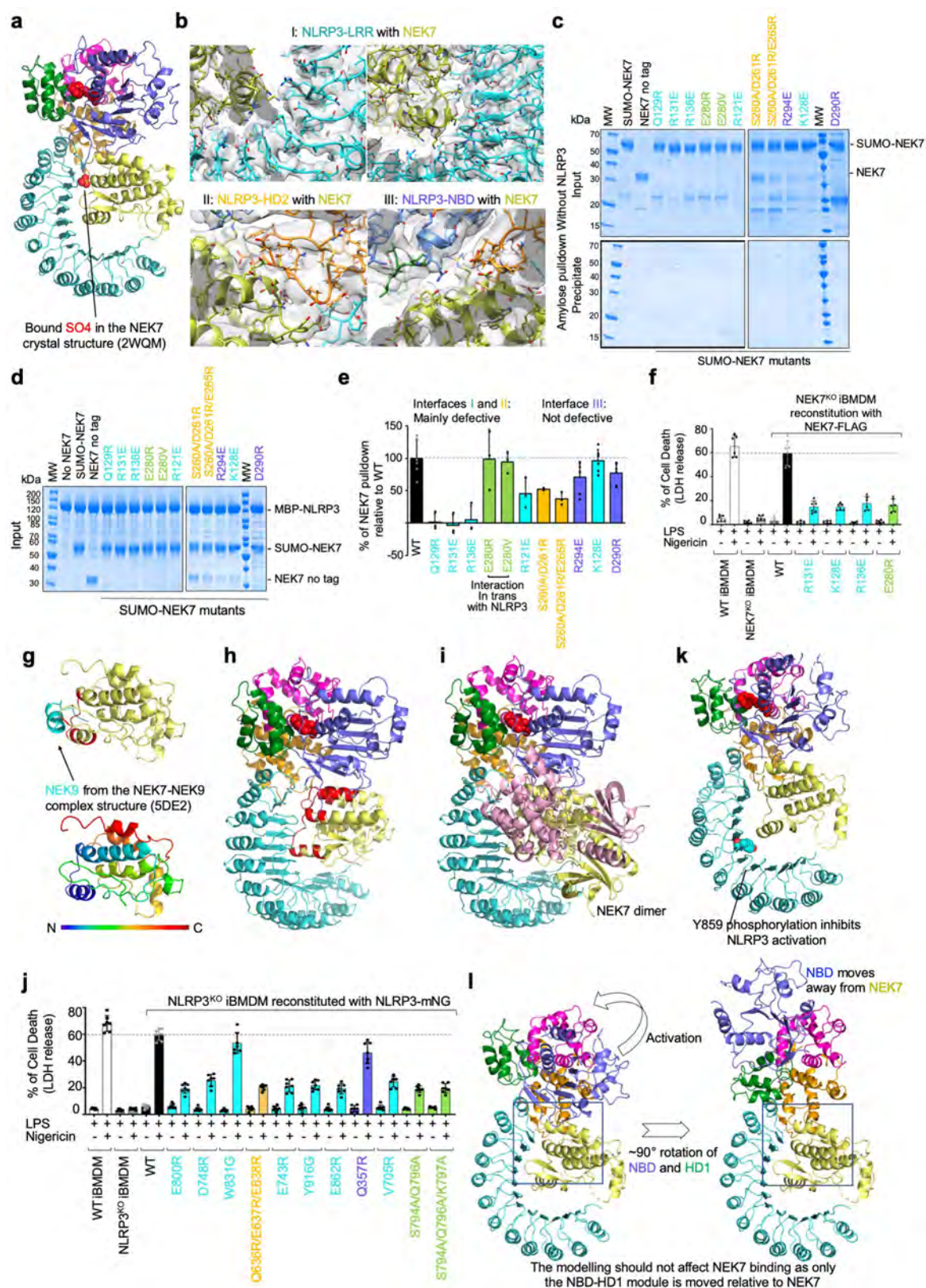
interaction. **c,** NBD, HD1, WHD and HD2 of NLRP3, NOD2 (PDB code 5IRL) and NLRC4 (PDB code 4KXF) superimposed and shown side-by-side for comparison. **d, e,** NOD2 (**d**) and NLRC4 (**e**) structures^{6,21} in inactive conformations and in the orientation of NLRP3 (shown in **a**) by superposing on the NBD and HD1 domains. The location of phosphorylated S533 of NLRC4 is labelled.



Extended Data Fig. 6 | Multiple sequence alignment of NEK7 and NEK6. Multiple sequence alignment was performed by the eScript server⁵⁴. Annotations are based on the PISA server²⁸ analysis, and mutational data from Figs. 3, 4.



Extended Data Fig. 7 | Multiple sequence alignment of NLRP3. Multiple sequence alignment was performed by the eScript server⁵⁴. Annotations are based on the PISA server²⁸ analysis, and mutational data from Figs. 3, 4.

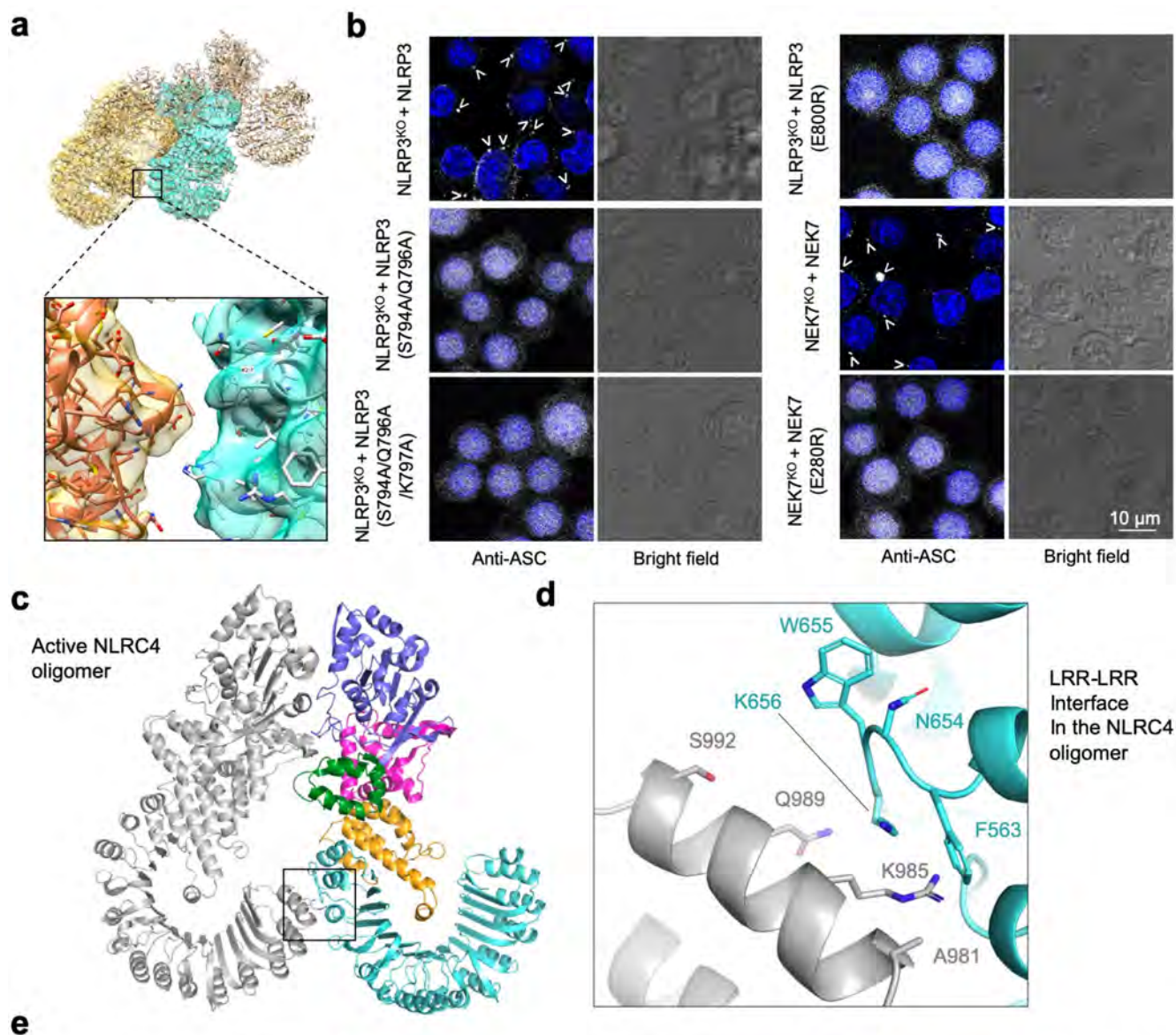


Extended Data Fig. 8 | See next page for caption.

Extended Data Fig. 8 | Structural analysis, mutagenesis tests and overlap between the NEK9- and NLRP3-binding sites on NEK7.

a, Position of the bound SO_4 ion as in the NEK7 crystal structure²⁵, superimposed here in the structure of the NLRP3–NEK7 complex. **b**, Cryo-EM densities (5σ) at the NLRP3–NEK7 interfaces. Views correspond to those of Fig. 3b, c. **c**, Amylose pulldown of wild-type and mutant NEK7 without NLRP3, showing input and precipitate gels, which serve as negative controls for Fig. 3d. Experiments were repeated 3–6 times. **d**, Input gel for amylose pulldown of wild-type and mutant NEK7 shown in Fig. 3d. Experiments were repeated 3–6 times. **e**, Alternative calculation of the percentage of NEK7 pulldown relative to wild type, by subtracting the NEK7/NLRP3 ratio in the absence of NEK7 from the observed NEK7/NLRP3 ratio (shown as mean \pm s.d. for $n = 3$ –6 experiments). **f**, Cytoplasmic LDH released into the supernatant, quantified by comparison with total intracellular LDH of untreated cells (triton-X-lysed). Data are presented as mean \pm s.d. for $n = 3$ replicates from 2 independent experiments. **g**, Top, mapping the NEK9-binding site onto the structure of the NEK7–NEK9 complex (PDB code 5DE2)²⁹. NEK7 and NEK9 are shown in yellow and cyan, respectively, with the

NEK9-binding residues of NEK7 highlighted in red. Bottom, a rainbow-coloured NEK7 structure showing that the NEK9-binding residues are from the first part of the NEK7 C-lobe. **h**, Mapping the NLRP3-binding site (red) onto NEK7 in the NLRP3–NEK7 structure. The NLRP3-binding site overlaps with the NEK9-binding site on NEK7. **i**, Superposition of NEK7 back-to-back dimer (yellow and pink) in the NEK7–NEK9 complex structure (PDB code 5DE2)²⁹ onto the NEK7 monomer in the NLRP3–NEK7 structure. The pink monomer in the NEK7 dimer clashes with NLRP3, which suggests that NEK7 dimerization cannot occur in the NLRP3–NEK7 complex. **j**, NLRP3-knockout iBMDMs were reconstituted with wild-type or mutant human mNG-tagged NLRP3, primed by LPS (4 h) and stimulated by nigericin (30 min). LDH release was analysed as in **f**. Dots, individual data points. Data are presented as mean \pm s.d. for $n = 3$ replicates from 2 independent experiments. **k**, LRR phosphorylation at Y859³⁰, which might cause steric and charge repulsion with NEK7. **l**, NLRP3 activation model and NEK7 interactions. The hypothetical $\sim 90^\circ$ rotation of the NBD and HD1 of NLRP3 upon activation moves the NBD away from the NEK7 interaction region (indicated by a box).



Extended Data Fig. 9 | See next page for caption.

Extended Data Fig. 9 | NLRP3–NEK7 interaction *in trans* and NLRP3 CAPS mutations. **a**, Cryo-EM density (5σ) fitted to the oligomeric model, zoomed in at an NLRP3–NEK7 interface *in trans*. **b**, Immunofluorescence imaging for assessing ASC speck formation. NLRP3-knockout and NEK7-knockout iBMDMs were reconstituted with wild-type and mutant constructs as depicted. ASC speck formation upon nigericin activation was analysed by anti-ASC antibody. Nucleus was stained by Hoechst 33342. Scale bar, 10 μ m. Images are representative of two independent experiments. **c**, An active NLRC4 oligomer structure²⁰ shown for the LRR–LRR interaction. **d**, Magnified view showing the detailed LRR–LRR

interaction *in trans* boxed in **c**. **e**, Mapping of pathogenic disease mutations onto the NLRP3 structure. CAPS-associated pathogenic mutations derived from the Infevers database³³ are shown with their predicted effect, based on the NLRP3 structure. Green, disruption of inter-domain interactions in the inactive conformation; mauve, change of local conformation by mutating residues buried within the domains; yellow, alteration in key residues in the Walker B motif; grey, enhancement of NEK7 binding. CINCA, chronic infantile neurological cutaneous articular; NOMID, neonatal-onset multisystem inflammatory disorder; FCAS, familial cold auto-inflammatory syndrome.

Extended Data Table 1 | Cryo-EM data collection, refinement and validation statistics

	NLRP3-NEK7 Talos Arctica dataset	NLRP3-NEK7 Titan Krios dataset
Data collection and processing		
Magnification	105,000	165,000
Voltage (keV)	200	300
Electron exposure (e-/Å ²)	~59	~55
Defocus range (μm)	-0.5 to -3.0	-0.8 to -3.0
Super Resolution pixel size (Å)	0.589	0.42
Symmetry imposed	C1	C1
Initial particle images (no.)	1,086,055	1,850,332
Final particle images (no.)	98,340	108,771
Map resolution (Å)	4.3	3.8
FSC threshold	0.143	0.143
Map resolution range (Å)	3.5-5.8	2.5-6.6
Refinement		
Initial model used		
Model resolution (Å)		4.3
FSC threshold		0.5
Model resolution range (Å)		2.5-6.6
Map sharpening B factor (Å ²)		-100
Model composition		
Non-hydrogen atoms		7979
Protein residues		980
Ligands		1
B factors (Å ²)		
Protein		132.7
Ligand		96.7
R.m.s. deviations		
Bond lengths (Å)		0.006
Bond angles (°)		1.345
Validation		
MolProbity score		2.35
Clashscore		8.92
Poor rotamers (%)		1.91
Ramachandran plot		
Favoured (%)		84.96
Allowed (%)		14.02
Disallowed (%)		1.02

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a Confirmed

- ☐ ☒ The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- ☐ ☒ An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- ☐ ☒ The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- ☒ ☐ A description of all covariates tested
- ☐ ☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- ☒ ☐ A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- ☒ ☐ For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
Give P values as exact values whenever suitable.
- ☒ ☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- ☒ ☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- ☒ ☐ Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated
- ☐ ☒ Clearly defined error bars
State explicitly what error bars represent (e.g. SD, SE, CI)

Our web collection on [statistics for biologists](#) may be useful.

Software and code

Policy information about [availability of computer code](#)

Data collection

Cryo-EM data was carried out with SerialEM 3.7

Data analysis

Microscale Thermophoresis data analysis was done with MO Control software provided by (NanoTemper, Munich, Germany)
Multi-angle light scattering (MALS) data analysis was carried out using ASTRA V
Cryo-EM data analysis was done with MotionCor2, CTFFIND4, RELION 3.0, cisTEM 1.0.0., Gautomatch v0.56, Gctf v1.18, ROME 1.1.2 package, Coot-0.8.9.2, Chimera1.13.1, Pymol 2.3.1
Phyre2 server <http://www.sbg.bio.ic.ac.uk/phyre2/html/page.cgi?id=index>
Thermal Shift assay data was analyzed with Protein Thermal ShiftTM Software (Thermo Fisher Scientific)
Adobe Photoshop software for image processing
Statistical analysis were performed using MS Excel and GraphPad Prism 7
All references are given in the Material and Methods section

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The cryo-EM map has been deposited in the Electron Microscopy Data Bank under the accession number EMD-0476. The atomic coordinates have been deposited in the Protein Data Bank under the accession number 6NPY.

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/authors/policies/ReportingSummary-flat.pdf](https://www.nature.com/authors/policies/ReportingSummary-flat.pdf)

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No statistical methods were used to predetermine sample sizes.
Data exclusions	No data were excluded.
Replication	All experiments were confirmed with multiple biological replicates as detailed in Methods or Figure Legends
Randomization	No randomization was performed.
Blinding	No blinding is used.

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
<input type="checkbox"/>	<input checked="" type="checkbox"/> Unique biological materials
<input type="checkbox"/>	<input checked="" type="checkbox"/> Antibodies
<input type="checkbox"/>	<input checked="" type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Human research participants

Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

Unique biological materials

Policy information about [availability of materials](#)

Obtaining unique materials All data can be obtained from the corresponding author upon reasonable request.

Antibodies

Antibodies used

anti-human NLRP3 antibody (Adipogen, Cat no: AG-20B-0014-C100), anti-mouse caspase-1 antibody (Adipogen, Cat. no: AG-20B-0042-C100), anti-FLAG F1804-Sigma, Anti-NEK7 antibody [EPR4900] Abcam, anti-ASC Cell Signaling Technology, Cat. no: 67824S, Goat anti-rabbit (H+L) Alexa Fluor 647 conjugated, Thermo Fischer Scientific, Cat. no: A- 21245, Anti-β-actin (mouse monoclonal, Santa Cruz Biotechnology, Cat. no: 47778)

Eukaryotic cell lines

Policy information about [cell lines](#)

Cell line source(s)

NLRP3 KO iBMDM was a kind gift by Kate Fitzgerald and NEK7 KO iBMDM by Gabreil Nunez

HEK293T (ATCC) <https://www.atcc.org/en/Products/All/CRL-3216.aspx>

SF9 cells from Thermo Fischer Scientific <https://www.thermofisher.com/order/catalog/product/11496015>

BL21(DE3) from Agilent [https://www.agilent.com/en/product/protein-expression/competent-cells-for-routine-protein-expression/general-protein-expression/bl21\(de3\)-competent-cells-232943](https://www.agilent.com/en/product/protein-expression/competent-cells-for-routine-protein-expression/general-protein-expression/bl21(de3)-competent-cells-232943)

Authentication

Cell lines were verified by manufacturer's website and Identity of these cell lines were frequently checked by their morphological features

HEK293T (ATCC) <https://www.atcc.org/en/Products/All/CRL-3216.aspx>

SF9 cells from Thermo Fischer Scientific <https://www.thermofisher.com/order/catalog/product/11496015>

BL21(DE3) from Agilent [https://www.agilent.com/en/product/protein-expression/competent-cells-for-routine-protein-expression/general-protein-expression/bl21\(de3\)-competent-cells-232943](https://www.agilent.com/en/product/protein-expression/competent-cells-for-routine-protein-expression/general-protein-expression/bl21(de3)-competent-cells-232943)

Mycoplasma contamination

All cell lines were tested to be mycoplasma-negative by PCR.

Commonly misidentified lines (See [ICLAC](#) register)

No commonly misidentified cell lines are used in this study.

Perfect Andreev reflection due to the Klein paradox in a topological superconducting state

Seunghun Lee^{1,2}, Valentin Stanev^{1,2,3}, Xiaohang Zhang^{1,2}, Drew Stasak¹, Jack Flowers¹, Joshua S. Higgins^{2,4}, Sheng Dai⁵, Thomas Blum⁶, Xiaoqing Pan^{5,6,7}, Victor M. Yakovenko^{3,4,8}, Johnpierre Paglione^{2,4}, Richard L. Greene^{2,4}, Victor Galitski^{3,4,8} & Ichiro Takeuchi^{1,2,4,*}

In 1928, Dirac proposed a wave equation to describe relativistic electrons¹. Shortly afterwards, Klein solved a simple potential step problem for the Dirac equation and encountered an apparent paradox: the potential barrier becomes transparent when its height is larger than the electron energy. For massless particles, backscattering is completely forbidden in Klein tunnelling, leading to perfect transmission through any potential barrier^{2,3}. The recent advent of condensed-matter systems with Dirac-like excitations, such as graphene and topological insulators, has opened up the possibility of observing Klein tunnelling experimentally^{4–6}. In the surface states of topological insulators, fermions are bound by spin–momentum locking and are thus immune from backscattering, which is prohibited by time-reversal symmetry. Here we report the observation of perfect Andreev reflection in point-contact spectroscopy—a clear signature of Klein tunnelling and a manifestation of the underlying ‘relativistic’ physics of a proximity-induced superconducting state in a topological Kondo insulator. Our findings shed light on a previously overlooked aspect of topological superconductivity and can serve as the basis for a unique family of spintronic and superconducting devices, the interface transport phenomena of which are completely governed by their helical topological states.

Klein’s gedanken experiment illustrates the intrinsic connection between particles and antiparticles in relativistic quantum mechanics, and observing this connection ostensibly requires velocities close to the speed of light². However, several condensed-matter systems have recently emerged as unexpected platforms for the study of relativistic effects. In materials such as graphene and topological insulators, the Dirac equation provides an effective low-energy description of band electrons^{4,5}. In graphene heterostructures, the modulation of conductance as functions of electron trajectory and electrostatic potential profile has previously been used as a vehicle for the investigation of Klein tunnelling^{5,7,8}. Here we demonstrate an alternative way in which to directly observe Klein tunnelling using a topological insulator. We use point-contact Andreev reflection (PCAR) measurements at the interface between a normal metal and a topological superconducting state (that is, the superconducting surface states of a topological insulator). The perfect transmission of electrons through a finite barrier manifests as an observed doubling of the conductance within the superconducting gap (Δ). This doubling of the conductance is due to the conservation of charge, spin and momentum in the Andreev reflection process, which requires that a positively charged hole with opposite spin and momentum to that of the electron is left behind^{9–11}. In real experiments, however, enhancement of the conductance is easily suppressed by various inevitable scattering mechanisms that arise from non-ideal interface conditions, and complete doubling of the conductance is very rarely observed. The extreme sensitivity to scattering makes Andreev reflection a unique tool for the detection of Klein tunnelling.

Spin–momentum locking of the Dirac states prohibits the reflection of an incident electron normal to the interface, irrespective of the microscopic details of the interface¹². It results in the complete absence of backscattering, and thus gives rise to topologically protected perfect Andreev reflection that manifests as an exact doubling of the conductance. Such a direct probe for the observation of Dirac particles could lead to a better understanding of their condensed matter implementations, and greater use of their properties in quantum transport devices.

To investigate how the presence of Dirac states at the surface of a topological insulator affects the processes of particle transport governed by Andreev reflection, we used a tip made of a platinum-iridium alloy (PtIr) to form a point-contact interface with a topological-insulator film in which superconductivity is induced through the proximity effect (Fig. 1a). We used heterostructures consisting of samarium hexaboride (SmB₆) and yttrium hexaboride (YB₆) to induce superconductivity in the Dirac surface states of SmB₆. SmB₆ is a topological Kondo insulator, in which the bulk gap at low temperatures ensures the existence of an insulating bulk sandwiched by topologically protected conducting surface layers^{13–18}. This is a critical prerequisite for the observation of effects that originate solely from the topologically protected states^{19,20}. The use of the isostructural rare-earth-hexaboride superconductor YB₆ (with a critical temperature, T_c , of around 6.3 K) as the layer underneath SmB₆ enables the fabrication, by sequential high-temperature growth, of a pristine SmB₆/YB₆ interface, which is necessary for achieving a robust proximity effect²¹ (see Methods and Extended Data Figs. 1, 2 for details).

As theoretically predicted⁶ and experimentally confirmed²⁰, the superconducting proximity effect that occurs in such topological insulator/superconductor heterostructures creates helical Cooper pairing on the surface of a topological insulator. Owing to the constraints imposed by the two-dimensional surface states and the insulating bulk, incoming electrons with finite momenta perpendicular to the surface (p_z) do not participate in the transport at the interface between a normal metal and topological insulator/superconductor heterostructure²². Thus, the PtIr-SmB₆/YB₆ contact creates an interface at which only in-plane transport (that is, momentum parallel to the plane of the surface states, so $p_z = 0$) is allowed (Fig. 1a). In addition, induced spin–momentum locking in a normal metal in contact with a topological insulator has previously been observed as a result of the topological proximity effect^{23,24}. Owing to the spin–momentum locking on both sides, incident electrons are forbidden from reflecting back (Fig. 1b). The perfect electron transmission to superconducting SmB₆ and the concomitant hole generation result in the observed doubling of conductance for energies within the proximity-induced Δ .

SmB₆/YB₆ heterostructures were analysed by point-contact spectroscopy at 2 K. For SmB₆ layers with thicknesses in the range of 20 to 30 nm, normalized differential-conductance (dI/dV) curves showed doubling of the conductance within the bias voltage corresponding

¹Department of Materials Science and Engineering, University of Maryland, College Park, MD, USA. ²Center for Nanophysics and Advanced Materials, University of Maryland, College Park, MD, USA. ³Joint Quantum Institute, University of Maryland, College Park, MD, USA. ⁴Department of Physics, University of Maryland, College Park, MD, USA. ⁵Department of Materials Science and Engineering, University of California, Irvine, CA, USA. ⁶Department of Physics and Astronomy, University of California, Irvine, CA, USA. ⁷Irvine Materials Research Institute, University of California, Irvine, CA, USA. ⁸Condensed Matter Theory Center, University of Maryland, College Park, MD, USA. *e-mail: takeuchi@umd.edu

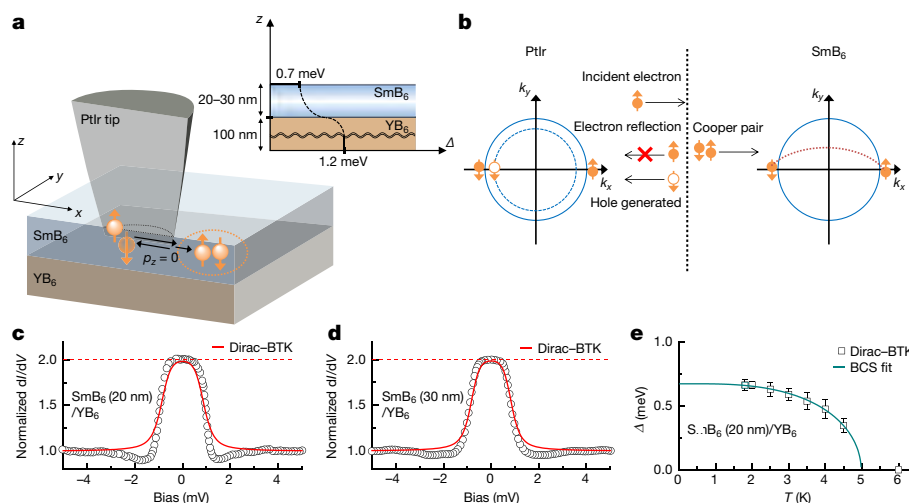


Fig. 1 | Perfect Andreev reflection due to the Klein paradox.

a, Schematic of PCAR measurement on SmB₆/YB₆ heterostructures. Owing to the lack of bulk states in SmB₆, only electrons with momentum parallel to the plane of the surface states of SmB₆ (that is, $p_z = 0$) contribute to transport. The inset shows variation of Δ in a SmB₆ (20–30 nm)/YB₆ heterostructure. There is a finite Δ in the top conducting surface of SmB₆. **b**, Andreev reflection process at the interface between PtIr and superconducting SmB₆. The surface of SmB₆ has topologically protected helical states exhibiting spin–momentum locking. Irrespective of barrier height, normal electron reflection is not allowed because it requires a spin flip. **c**, **d**, Perfect Andreev reflection due to Klein tunnelling, indicated

by exact doubling of the normalized differential conductance (dI/dV), is observed in the point-contact spectroscopy of PtIr–SmB₆ (20 nm)/YB₆ (100 nm) (**c**) and PtIr–SmB₆ (30 nm)/YB₆ (100 nm) (**d**) heterostructures measured at 2 K. The red lines are fits to the experimental data using a BTK model modified with a Dirac Hamiltonian (Dirac–BTK) with $\Delta = 0.75 \pm 0.06$ meV (**c**) and $\Delta = 0.73 \pm 0.05$ meV (**d**). **e**, The temperature-dependent Δ (extracted using the Dirac–BTK model) from a Au–SmB₆ (20 nm)/YB₆ structure in which a gold thin film was used to form the junction (see Methods, Extended Data Fig. 3), displaying Bardeen–Cooper–Schrieffer behaviour (cyan line).

to the induced Δ . As seen in Figs. 1c, 1d, the observed doubling of the conductance is exact within the uncertainty due to the fitting procedure (see Methods and Extended Data Fig. 5). In this regime, the SmB₆ layer is sufficiently thick to have fully developed topologically protected surface states, while the superconducting proximity effect from the YB₆ can still be observed at the top surface (as depicted in the inset of Fig. 1a).

The best theoretical fit to the data is based on the Blonder, Tinkham and Klapwijk (BTK) theory⁹ (see below) and results in a proximity-induced Δ of around 0.7 meV; as expected, this is smaller than the bulk Δ of YB₆ (≈ 1.3 meV)²⁵. The temperature dependence and magnetic-field dependence of a dI/dV spectrum were measured on a separately fabricated Au–SmB₆/YB₆ structure (see Methods and Extended Data Figs. 3, 7), in which the junction comprises a thin film of gold. The obtained temperature dependence of Δ shows the Bardeen–Cooper–Schrieffer behaviour, as expected (Fig. 1e); this confirms that the enhancement of conductance seen in the dI/dV spectrum is due to the proximity-induced superconductivity on the topologically protected top surface of SmB₆.

The transmission and reflection of particles through an interface between a normal metal and a superconductor is described by the BTK theory⁹. A dimensionless parameter Z represents the interfacial barrier strength, which reduces the transparency of the interface: perfect conductance doubling within Δ thus requires $Z \approx 0$ in the standard BTK theory. However, we show below that the superconductivity induced in the topologically protected surface states of SmB₆ inhibits electron reflection and cancels the effect of the barrier strength. Thus, even for a finite Z in PtIr–SmB₆/YB₆ point contact there can be perfect electron transmission, which is directly discernible by the conductance doubling. Describing this phenomenon requires modification of the BTK theory to account for the role of the Dirac surface states of SmB₆ in the dI/dV spectrum, which is henceforth referred to as the Dirac–BTK theory.

The perfect Andreev reflection is a direct consequence of the presence of topologically protected surface states as well as the absence of a bulk conduction channel. On the basis of our systematic study, when the thickness of the SmB₆ film is less than about 20 nm^{21,26} the

effect of hybridization of the top and bottom surface states becomes pronounced, which opens a gap in the dispersion of the surface states and weakens the topological protection^{20,27} (Fig. 2a). This accounts for the reduced conductance enhancement that is observed upon contact with the SmB₆ (10 nm)/YB₆ heterostructure at 2 K (Fig. 2b). To confirm the role of the robust bulk gap of SmB₆ in our observation, we also performed PCAR measurements on Sm_{1-x}Y_xB₆ (20 nm)/YB₆ heterostructures, in which samarium is partially substituted by yttrium in the top layer to modify its electronic structure. Yttrium ions are expected to generate conducting bulk states, which in turn give rise to transport channels that are not subjected to spin–momentum locking (see Methods and Extended Data Fig. 4). As expected, point-contact spectra of Sm_{0.8}Y_{0.2}B₆/YB₆ and Sm_{0.5}Y_{0.5}B₆/YB₆ heterostructures at 2 K show a conductance enhancement at zero bias of around 1.5, which is substantially lower than an exact doubling (Fig. 2c, d).

When the surface of a YB₆ film is probed directly—that is, with no SmB₆ layer on top—the point-contact spectrum at 2 K displays an entirely different characteristic: the junction is now in the regime in which tunnelling has a substantial contribution, resulting in reduced conductance in the gap region of YB₆ with a substantial barrier strength at the interface ($Z \approx 1$, extracted using the standard BTK model) (Fig. 2e). The gap value ($\Delta \approx 1.3$ meV) determined from the fit is consistent with the full superconducting gap of YB₆²⁵. At the other limit, when the thickness of SmB₆ is greater than 40 nm, the dI/dV spectrum at 2 K (Fig. 2f) does not show any features corresponding to proximity-induced superconductivity. Instead, the entire dI/dV spectrum shows Fano resonance—a familiar signature of the Kondo lattice physics of bulk SmB₆¹⁶.

To illustrate the uniqueness of the perfect Andreev reflection observed here, we surveyed the open literature on PCAR measurements performed on various superconductors. Figure 3 shows plots of normalized dI/dV at zero bias (that is, conductance enhancement) against Z (obtained from the BTK fit) from 44 reports selected from 250 publications on PCAR measurements (see Methods; the list of publications and other details are provided in Supplementary Table). The general trend is well captured by the standard BTK model (cyan line). To the best of our knowledge there are only two studies in the literature,

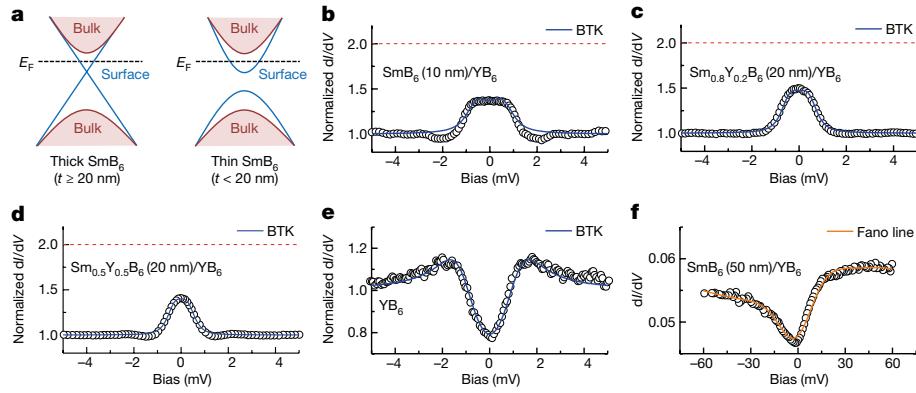


Fig. 2 | Sensitivity of perfect Andreev reflection to compromised topological superconductivity. When superconductivity in the surface states of SmB_6/YB_6 is modified by changing the thickness or the composition of the SmB_6 layer, the conductance doubling is suppressed. **a**, Band structures of different thicknesses of SmB_6 . **b**, Point-contact spectrum of a SmB_6 (10 nm)/ YB_6 heterostructure. Reduced conductance at zero bias (normalized $dI/dV \approx 1.4$) is observed. **c–e**, Point-contact spectra of yttrium-substituted SmB_6 ($\text{Sm}_{1-x}\text{Y}_x\text{B}_6$ (20 nm))/ YB_6 heterostructures with $x = 0.2$ (**c**) and $x = 0.5$ (**d**) and of the YB_6 layer only (**e**). The blue lines are best fits to the standard BTK theory: for **b**, $Z = 0.42 \pm 0.10$,

$\Delta = 0.59 \pm 0.10$ meV and $\Gamma \leq 0.16$ meV; for **c**, $Z = 0.35 \pm 0.09$, $\Delta = 0.49 \pm 0.05$ meV and $\Gamma \leq 0.08$ meV; for **d**, $Z = 0.42 \pm 0.06$, $\Delta = 0.30 \pm 0.04$ meV and $\Gamma \leq 0.04$ meV; for **e**, $Z = 1.04 \pm 0.06$, $\Delta = 1.24 \pm 0.08$ meV and $\Gamma = 0.60 \pm 0.04$ meV. Z and Γ are the interface barrier strength and the broadening parameter, respectively. **f**, The point-contact spectrum of a SmB_6 (50 nm)/ YB_6 exhibits an asymmetric Fano-like spectrum due to the inherent Kondo-lattice electronic structure of SmB_6 . The orange line is the best fit to the Fano-line shape¹⁶. All point-contact spectra were obtained at 2 K.

both on Nb–Cu junctions^{28,29}, that report an observed normalized dI/dV at zero bias of greater than 1.9. A detailed explanation and comparison between our PtIr– SmB_6 (20–30 nm)/ YB_6 point-contact spectra and the reported Nb–Cu point-contact spectra (Extended Data Fig. 8) are provided in Methods.

According to the standard BTK theory, the observed perfect conductance doubling implies that, when the thicknesses of the SmB_6 layer are in the range of 20 to 30 nm, $Z \approx 0$ for contacts to the SmB_6/YB_6 heterostructures. However, we think that the factors dictated by the materials are similar or identical for all heterostructures studied here, including the 10-nm-thick SmB_6 and the yttrium-substituted SmB_6 heterostructures. Thus, we expect materials-dictated Z for junctions that exhibit perfect Andreev reflection to be approximately 0.4—an average of the extracted Z values for the contacts with heterostructures that do not have complete topological protection (navy pentagon in Fig. 3).

Now we consider the mechanism of the perfect Andreev reflection for finite Z . To describe the transmission and the reflection processes at an interface between a normal metal and a superconducting topological insulator (which in our case is a topological insulator with proximity-induced superconductivity in the surface states), we modify the standard BTK theory⁹—which describes the transport at a normal metal/conventional superconductor interface—by considering the unique properties of a superconducting topological insulator. The key factor in the modification is the interplay of the spin and the momentum of the electrons in the surface states of SmB_6 —a consequence of the non-trivial topology of the bulk band structure. These states are described by the Dirac Hamiltonian that displays spin–momentum locking, as manifested in helicity. As first shown by Klein, this can lead to perfect transmission through an arbitrarily large potential barrier: normal reflection of a Dirac particle requires a complete spin flip and thus is forbidden. The presence of such perfectly transmitting channels at the boundary between a topological material and a topological superconductor nullifies the effects of the boundary barrier, including the Fermi velocity mismatch, thus leading to perfect Andreev reflection—that is, the doubling of the conductance within Δ in a dI/dV spectrum¹². Bulk PtIr is a normal metal. However, the topological proximity effect can render PtIr topologically nontrivial when in contact with SmB_6 , thereby satisfying the necessary condition for the perfect Andreev reflection^{23,24}: the contact with the SmB_6 surface breaks the degeneracy of the two helicities in the PtIr tip, and in the region adjacent to the interface only the states matching the helicity on the SmB_6 side are allowed. The strong spin–orbit coupling

of PtIr itself can also play a part in this process (see Supplementary Discussion for details).

We thus model the PtIr– SmB_6 boundary as a line dividing the normal and the superconducting regions in the plane of the SmB_6 surface states. At the boundary, we add a delta-function potential term $U(x) = U_0\delta(x)$ modelling the barrier at the interface, typically represented by the dimensionless barrier-strength parameter Z ($Z \equiv U_0/\hbar v_F^S$), where \hbar is the reduced Planck's constant and v_F^S is the Fermi velocity on the SmB_6 side. The Dirac Hamiltonian on the superconducting topological-insulator side can be written (in $\Psi = [\psi_{\uparrow,\varepsilon,\mathbf{p}}, \psi_{\downarrow,\varepsilon,\mathbf{p}}, \psi_{\uparrow,-\varepsilon,-\mathbf{p}}, \psi_{\downarrow,-\varepsilon,-\mathbf{p}}]$ basis) as³⁰

$$H_{\text{hetero}} = \begin{pmatrix} v_F \mathbf{p} \cdot \boldsymbol{\sigma} - \sigma_0 \mu + \sigma_0 U(x) & i\sigma_y \Delta \\ -i\sigma_y \Delta & v_F \mathbf{p} \cdot \boldsymbol{\sigma}^* + \sigma_0 \mu - \sigma_0 U(x) \end{pmatrix}$$

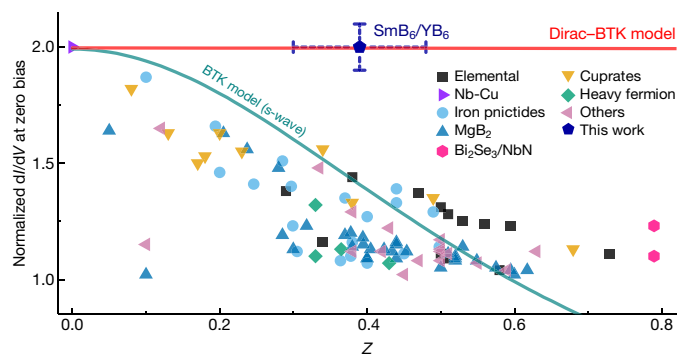


Fig. 3 | Andreev reflection process under the Dirac Hamiltonian.

A survey of reported normalized dI/dV at zero bias (that is, conductance enhancement) plotted against Z (that is, the dimensionless barrier strength parameter) for PCAR measurements on various superconductors (see Supplementary Table for references and plotted values). The theoretical Z -dependent normalized dI/dV at zero bias, calculated by the standard BTK (cyan) and the Dirac–BTK models (red) are shown (see Methods and Extended Data Fig. 6 for details). Simulation parameters: $T = 2$ K; $\Delta = 1$ meV. For PtIr– SmB_6 (20–30 nm)/ YB_6 junctions that display perfect Andreev reflection (normalized $dI/dV = 2$), we assume that the Z value is similar to that of junctions with other heterostructures in this study that do not have perfect Andreev reflection (thin SmB_6 (10 nm) and Y-substituted SmB_6); that is, $Z = 0.39 \pm 0.09$, with error bars reflecting the fitting procedure (navy pentagon, this work).

where \mathbf{p} is momentum in the x - y plane, μ is the chemical potential and $\sigma \equiv [\sigma_x, \sigma_y]$ ($\{\sigma_0, \sigma_x, \sigma_y, \sigma_z\}$ is the set of the identity and the Pauli matrices in the spin space). Δ is the proximity-induced superconducting gap in the top surface of the SmB_6 layer.

Using the appropriate boundary condition for a metal with single helicity (see Supplementary Discussion for details) and for energies close to the Fermi level, we can derive analytically the coefficients for each allowed process: r_e , reflection as an electron; r_h , Andreev reflection; t_e , transmission as an electron-like particle; t_h , transmission as a hole-like particle. These coefficients depend on the following: the energy (or bias voltage V); θ_k , the angle of incidence measured from the normal to the boundary; Z , which encodes the effects of the boundary barrier; and v_F^S/v_F^N , the Fermi velocity mismatch, where v_F^N is the Fermi velocity on the normal metal (PtIr) side.

The conductance ($G = dI/dV$) through the interface is then given (at zero temperature) by:

$$G = \frac{dI}{dV} = G_0 \int_{-\chi}^{\chi} (1 - |r_e(\theta_k)|^2 + |r_h(\theta_k)|^2) f_{\theta_k} \cos \theta_k d\theta_k \quad (1)$$

where f_{θ_k} models the angular distribution of the incoming electrons, $\chi \equiv \arcsin(v_F^N/v_F^S)$, and G_0 is a constant. The angular dependence of r_e goes as $r_e(\theta_k) \approx \sin \theta_k$; reflection as an electron at $\theta_k = 0$ requires a spin flip, which is forbidden by time-reversal symmetry, and thus $r_e(\theta_k = 0) = 0$. Reproducing the observed perfect conductance doubling requires a rather narrow f_{θ_k} centred around $\theta_k = 0$ (see Supplementary Discussion for details). In this quasi-one-dimensional case, there is perfect transmission irrespective of the barrier height and Fermi velocity mismatch—the essence of Klein tunnelling (red line in Fig. 3). For $|eV| < \Delta$ (that is, energies below the superconducting gap) this leads to $r_h(\theta_k \approx 0) = 1$, whereas for $|eV| \gg \Delta$ we have $r_h = 0$; combining these two results with equation (1) immediately leads to conductance doubling: $G(|eV| < \Delta)/G(|eV| \gg \Delta) = 2$.

In summary, we have observed perfect Andreev reflection—a manifestation of Klein tunnelling—using proximity-induced superconductivity in a three-dimensional topological insulator. Despite the formal similarity between Dirac excitations in graphene and in topological insulators, there are important differences between the two with respect to Klein tunnelling. In graphene, the degeneracy between sublattices of the honeycomb structure is crucial, whereas in topological insulators it is the time-reversal symmetry that directly prohibits backscattering. The unusual combination of the topologically protected surface states and the lack of bulk states in thin layers of SmB_6 films has facilitated the observation of perfect Andreev reflection due to Klein tunnelling. Perfect transmission renders transport of individual electrons across an interface dissipation-less, regardless of the origins of the potential barrier and its variation—an attractive attribute for many device applications including quantum information processing³¹ and high-sensitivity detectors³². We foresee Klein tunnelling in topological insulators to be a platform for the exploration of various interface transport phenomena, including perfect spin-filters as governed by unadulterated spin-momentum locking³³.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-019-1305-1>.

Received: 25 June 2018; Accepted: 16 April 2019;
Published online 19 June 2019.

- Dirac, P. A. M. The quantum theory of the electron. *Proc. R. Soc. Lond. A* **117**, 610–624 (1928).
- Klein, O. Die reflexion von Elektronen an einem Potentialsprung nach der relativistischen Dynamik von Dirac. *Z. Phys.* **53**, 157–165 (1929).
- Calogeracos, A. & Dombey, N. History and physics of the Klein paradox. *Contemp. Phys.* **40**, 313–321 (1999).
- Hasan, M. Z. & Kane, C. L. Colloquium: topological insulators. *Rev. Mod. Phys.* **82**, 3045–3067 (2010).

- Beenakker, C. W. J. Colloquium: Andreev reflection and Klein tunneling in graphene. *Rev. Mod. Phys.* **80**, 1337–1354 (2008).
- Fu, L. & Kane, C. L. Superconducting proximity effect and Majorana fermions at the surface of a topological insulator. *Phys. Rev. Lett.* **100**, 096407 (2008).
- Stander, N., Huard, B. & Goldhaber-Gordon, D. Evidence for Klein tunneling in graphene p-n junctions. *Phys. Rev. Lett.* **102**, 026807 (2009).
- Young, A. F. & Kim, P. Quantum interference and Klein tunneling in graphene heterojunctions. *Nat. Phys.* **5**, 222–226 (2009).
- Blonder, G. E., Tinkham, M. & Klapwijk, T. M. Transition from metallic to tunneling regimes in superconducting microconstrictions: Excess current, charge imbalance, and supercurrent conversion. *Phys. Rev. B* **25**, 4515–4532 (1982).
- Daghero, D. & Gonnelli, R. S. Probing multiband superconductivity by point-contact spectroscopy. *Supercond. Sci. Technol.* **23**, 043001 (2010).
- Lee, W.-C. & Greene, L. H. Recent progress of probing correlated electron states by point contact spectroscopy. *Rep. Prog. Phys.* **79**, 094502 (2016).
- Adroguer, P. et al. Probing the helical edge states of a topological insulator by Cooper-pair injection. *Phys. Rev. B* **82**, 081303 (2010).
- Dzero, M., Sun, K., Coleman, P. & Galitski, V. Theory of topological Kondo insulators. *Phys. Rev. B* **85**, 045130 (2012).
- Syers, P., Kim, D., Fuhrer, M. S. & Paglione, J. Tuning bulk and surface conduction in the proposed topological Kondo insulator SmB_6 . *Phys. Rev. Lett.* **114**, 096601 (2015).
- Yo, Y. S., Sun, K., Kurdak, Ç., Kim, D.-J. & Fisk, Z. Inverted resistance measurements as a method for characterizing the bulk and surface conductivities of three-dimensional topological insulators. *Phys. Rev. Appl.* **9**, 044006 (2018).
- Zhang, X. et al. Hybridization, inter-ion correlation, and surface states in the Kondo insulator SmB_6 . *Phys. Rev. X* **3**, 011011 (2013).
- Jiang, J. et al. Observation of possible topological in-gap surface states in the Kondo insulator SmB_6 by photoemission. *Nat. Commun.* **4**, 3010 (2013).
- Neupane, M. et al. Surface electronic structure of the topological Kondo-insulator candidate correlated electron system SmB_6 . *Nat. Commun.* **4**, 2991 (2013).
- Dai, W. et al. Proximity-effect-induced superconducting gap in topological surface states – a point contact spectroscopy study of $\text{NbSe}_2/\text{Bi}_2\text{Se}_3$ superconductor-topological insulator heterostructures. *Sci. Rep.* **7**, 7631 (2017).
- Xu, S.-Y. et al. Momentum-space imaging of Cooper pairing in a half-Dirac-gas topological superconductor. *Nat. Phys.* **10**, 943–950 (2014).
- Lee, S. et al. Observation of the superconducting proximity effect in the surface state of SmB_6 thin films. *Phys. Rev. X* **6**, 031031 (2016).
- Borisov, K., Chang, C.-Z., Moodera, J. S. & Stamenov, P. High Fermi-level spin polarization in the $(\text{Bi}_{1-x}\text{Sb}_x)_2\text{Te}_3$ family of topological insulators: a point contact Andreev reflection study. *Phys. Rev. B* **94**, 094415 (2016).
- Shoman, T. et al. Topological proximity effect in a topological insulator hybrid. *Nat. Commun.* **6**, 6547 (2015).
- Hutasoit, J. A. & Stanesco, T. D. Induced spin texture in semiconductor/topological insulator heterostructures. *Phys. Rev. B* **84**, 085103 (2011).
- Szabó, P. et al. Superconducting energy gap of YB_3 studied by point-contact spectroscopy. *Physica C* **460–462**, 626–627 (2007).
- Alexandrov, V., Coleman, P. & Erten, O. Kondo breakdown in topological Kondo insulators. *Phys. Rev. Lett.* **114**, 177202 (2015).
- Wang, M.-X. et al. The coexistence of superconductivity and topological order in the Bi_2Se_3 thin films. *Science* **336**, 52–55 (2012).
- Soulen Jr, R. J. et al. Measuring the spin polarization of a metal with a superconducting point contact. *Science* **282**, 85–88 (1998).
- Strijkers, G. J., Ji, Y., Yang, F. Y., Chien, C. L. & Byers, J. M. Andreev reflections at metal/superconductor point contacts: Measurement and analysis. *Phys. Rev. B* **63**, 104510 (2001).
- Tkachov, G. & Hankiewicz, E. M. Helical Andreev bound states and superconducting Klein tunneling in topological insulator Josephson junctions. *Phys. Rev. B* **88**, 075401 (2013).
- Janvier, C. et al. Coherent manipulation of Andreev states in superconducting atomic contacts. *Science* **349**, 1199–1202 (2015).
- Kornev, V. K., Kolotinskiy, N. V., Levochkina, A. Y. & Mukhanov, O. A. Critical current spread and thermal noise in Bi-SQUID cells and arrays. *IEEE Trans. Appl. Supercond.* **27**, 1601005 (2017).
- Zhang, C., Lu, H.-Z., Shen, S.-Q., Chen, Y. P. & Xiu, F. Towards the manipulation of topological states of matter: a perspective from electron transport. *Sci. Bull. (Beijing)* **63**, 580–594 (2018).

Acknowledgements We thank Y. S. Eo for discussions on the properties of SmB_6 , F. C. Wellstood for discussions on the possible applications of superconducting Klein tunnelling devices, and H. M. Iftikhar Jaim for assistance with X-ray measurements. This project was funded by ONR N00014-13-1-0635; ONR N00014-15-1-2222; AFOSR number FA9550-14-1-0332; NSF (DMR-1410665); and C-SPIN, one of six centers of STARnet, a Semiconductor Research Corporation (SRC) programme sponsored by MARCO and DARPA. We acknowledge support from the Maryland NanoCenter. J.P. acknowledges support from the Gordon and Betty Moore Foundation's EPIQS Initiative through grant number GBMF4419. V.G. was supported by DOE-BES (DESC0001911) and the Simons Foundation. This work was also supported in part by the Center for Spintronic Materials in Advanced Information Technologies (SMART), one of the centers in nCORE, an SRC programme sponsored by NSF and NIST. The work at University of California, Irvine, was carried out using the electron microscopy facilities

of the Irvine Materials Research Institute (IMRI) and was supported by the National Science Foundation through grant DMR-1506535 and by DOE-BES under grant DE-SC0014430. We acknowledge support from the National Institute of Standards and Technology Cooperative Agreement 70NANB17H301.

Reviewer information *Nature* thanks Ewelina Hankiewicz, David Goldhaber-Gordon and Jinfeng Jia for their contribution to the peer review of this work.

Author contributions S.L., X.Z. and I.T. conceived the experiment. S.L. fabricated thin films and devices, and performed their characterization—including point-contact spectroscopy measurements—with assistance from X.Z. and J.S.H. V.S., V.M.Y. and V.G. performed the theoretical calculations. D.S. analysed the compositions of the films. J.F. performed the literature survey on previous Andreev reflection experiments. S.D., T.B. and X.P. performed TEM measurements. V.M.Y., J.P., R.L.G. and V.G. helped with data interpretation and analysis and manuscript preparation. S.L., V.S., X.Z. and I.T. wrote the paper. I.T.

supervised and coordinated the project. All authors discussed the results and commented on the manuscript

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-019-1305-1>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-019-1305-1>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to I.T.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

METHODS

Fabrication of SmB₆ thin film. The growth conditions of SmB₆ thin films have been systematically optimized in order to ensure the quality of SmB₆ thin films. It is known that during the sputtering process, the considerable difference in the atomic masses of Sm and B leads to different scattering probabilities, and thus probably results in a B-deficient film when the deposition is carried out with a stoichiometric target^{21,34,35}. Therefore, we fabricated SmB₆ thin films on Si (001) substrates by co-sputtering SmB₆ and B targets to compensate for possible B deficiency. To remove the native oxide layer on the Si substrate, we treated with hydrofluoric acid (HF) before the thin film deposition. After reaching a base pressure of approximately 2×10^{-8} Torr, the sputtering process was performed on the Si substrates at 860 °C under a deposition pressure of 10 mTorr of Ar (99.999%). The distance between the targets and substrates, as well as the plasma density, were adjusted to increase the activation energy of sputtered species, which is correlated with chemical reaction and atomic migration³⁶. We optimized the power ratio of the two targets for the co-sputtering process by measuring the stoichiometry (that is, the B/Sm ratio) of the deposited SmB₆ thin films using wavelength dispersive spectroscopy (WDS). The optimal powers for SmB₆ and B were found to be 40 W and 60 W, respectively, for a distance between the targets and the substrate of about 10 cm. Under the optimized conditions, the B/Sm ratio of the SmB₆ thin film was 6.0 ± 0.1 . X-ray photoemission spectroscopy and energy-dispersive spectroscopy measurements of the films were used to verify the absence of any impurities that may give rise to metallic conduction at low temperatures. Temperature-dependent resistance measurements show the suggested signature of the emergence of metallic surface states—the saturation of the resistance at low temperatures (that is, resistance plateau) (see Extended Data Fig. 4).

Extended Data Fig. 1a shows a high-resolution transmission electron microscopy image of a cross-section of a SmB₆ sample. There is no indication of the presence of interfacial gradation or extra phases. Extended Data Fig. 1b–d shows selected area electron diffraction (SAED) patterns of the SmB₆ thin film, the Si substrate and the interface regions, respectively. The SAED pattern of the Si substrate (Extended Data Fig. 1c) shows the pattern along the [110] zone axis. In the SAED pattern of the interface (Extended Data Fig. 1d), an additional spot pattern corresponding to the SmB₆ [100] zone orientation (Extended Data Fig. 1b) can be clearly identified (indicated by yellow arrows). The result is indicative of the epitaxial relation, SmB₆ [100] || Si [110], which is consistent with a small lattice mismatch between Si (110) and SmB₆ (100) as illustrated in Extended Data Fig. 1e. Specifically, the *d*-spacing of Si (110) is 3.839 Å, and the lattice mismatch between Si (110) and SmB₆ (100) is about 7%. In addition, aberration-corrected scanning transmission electron microscopy was used, and the atomic-resolution image taken from the SmB₆ film (Extended Data Fig. 1f) reveals its cubic structure. The θ – 2θ X-ray diffraction pattern (Extended Data Fig. 1g) shows a *c*-axis-oriented structure of SmB₆. The XRD diffraction pattern exhibits sharp SmB₆ peaks, which are associated with the {001} planes only. The lattice parameter is found to be 4.13 Å, which is close to the bulk value¹⁴.

Fabrication of superconducting YB₆ thin films and the effect of stoichiometry on *T_c*. Yttrium hexaboride (YB₆) is a known rare-earth hexaboride superconductor with a bulk zero resistance *T_c* of around 7 K^{25,37}. It has been reported that the superconducting properties of YB₆ are closely related to the composition³⁸. However, a systematic study of the superconducting properties with broad variation in composition has not been previously reported. We have successfully fabricated superconducting YB_{6±δ} films for the first time. To achieve the highest *T_c* in YB_{6±δ} thin films for the present study, we first studied the effect of the stoichiometry on the *T_c* of sputtered YB_x thin films. Owing to the substantial difference in the atomic masses between Y and B, and the variation in the distance from the target to different locations of a 3" wafer, we were able to fabricate 'natural composition spread' films of YB_x by sputtering a stoichiometric YB₆ target. Similar to the deposition and the characterization of SmB₆ thin films, a deposition pressure of 10 mTorr and a growth temperature of 860 °C were used for YB_{6±x} thin film growth on Si (001) substrates. The distance between the YB₆ target and the Si substrate was about 10 cm, and the d.c. power applied to the YB₆ target was 60 W. The stoichiometric B/Y ratio for films deposited at different positions was examined by WDS measurements. As shown in Extended Data Fig. 2a, the temperature dependence of the normalized resistance (R/R_N , where *R_N* is the normal state resistance) of the YB_x thin films indicates that the superconducting transition temperature *T_c* varies with the stoichiometric B/Y ratio. In Extended Data Fig. 2b, *T_c* is plotted as a function of the stoichiometric B/Y ratio. The highest *T_c* is observed in the slightly boron-deficient region (B/Y = 5.6). Thus, YB_{5.6} films were used for the present study, and for simplicity, the YB_{5.6} films used in this study are referred to as YB₆ films. The SmB₆/YB₆ heterostructures were fabricated through a sequential high-temperature deposition process without breaking the vacuum—that is, an in situ process as described in the main text—to ensure a pristine interface between SmB₆ and YB_{5.6}²¹. YB₆ has a cubic structure with almost the same lattice constant as SmB₆ (about 4.1 Å) (YB₆: JCPDS number

16-0732 and SmB₆: JCPDS number 36-1326), and thus lattice mismatch strain is expected to be negligible.

Fabrication of Au-SmB₆/YB₆ structures and the temperature dependence of *dI/dV* curves. The analysis of the temperature dependence of *dI/dV* spectra can be used to verify that the gap-like feature is indeed attributed to the proximity-induced superconductivity. To perform a systematic temperature-dependence measurement, Au-SmB₆/YB₆ structures were fabricated using a method including multiple photolithography and ion-milling processes. Microposit S1813 was used as the photoresist, and after spin-coating the photoresist was baked at 100 °C for 2 min. After exposure to ultraviolet light, the samples were developed using a Microposit CD-26 developer for 60 s. A schematic cross-sectional structure of the thin-film devices is shown in Extended Data Fig. 3a. The SmB₆/YB₆ heterostructure was subject to in situ Ar plasma cleaning before Au deposition. After the Au deposition, two rounds of photolithography and ion-milling processes were carried out to define the line shape and the circular junction area, respectively. A SiO₂ (100 nm) layer was used to electrically isolate a top electrode from the SmB₆/YB₆ line. The top electrode, consisting of Au, was fabricated through a lift-off process. The optical microscope image (Extended Data Fig. 3b) shows the top view of a Au-SmB₆/YB₆ structure with a circular junction (diameter, 10 μm).

Extended Data Fig. 3c shows normalized *dI/dV* spectra of the Au-SmB₆ (20 nm)/YB₆ structure at different temperatures (1.8–4.5 K). The enhancement in conductance due to Andreev reflection is approximately 1.8, which is slightly smaller than the value obtained from junctions in the point-contact configuration as described in the main text. Given the specific geometric design of the junction, quasiparticle lifetime broadening^{39,40} and/or an oblique angle for incident electrons may lead to a slightly reduced zero-bias conductance enhancement (see Supplementary Discussion). In the former case, for example—as shown in Extended Data Fig. 3c—by introducing a lifetime broadening term Γ with a value of less than 10% of Δ , we can fit the data using the Dirac–BTK model. The Δ values obtained by the Dirac–BTK fits to the *dI/dV* spectra at different temperatures agree well with those from point-contact spectroscopy measurements carried out with a PtIr tip.

Comparison of SmB₆ and Y-substituted SmB₆. To confirm that the absence of bulk gapless states is crucial for the perfect conductance doubling observed in point-contact spectroscopy measurements, we modified the bulk electronic structure of SmB₆ by Y substitution. Specifically, we performed point-contact spectroscopy measurements on Sm_{1–x}Y_xB₆/YB₆ heterostructures. Y-substituted SmB₆ heterostructures were prepared by co-sputtering SmB₆, B and YB₆ targets, and the composition was determined by WDS. Extended Data Fig. 4a shows the resistance normalized by the value at 300 K (R/R_{300K} , logarithmic scale) plotted against the inverse of temperature ($1/T$) plots of SmB₆ as well as 20% and 50% Y-substituted SmB₆ (Sm_{0.8}Y_{0.2}B₆ and Sm_{0.5}Y_{0.5}B₆) thin films. The behaviour of the temperature-dependent resistance of the bulk states can be described by an exponential function, $R_{\text{bulk}}(T) \propto \exp(E_a/k_B T)$, where *E_a* and *k_B* are a carrier activation energy and Boltzmann constant, respectively. Hence the positive linear slopes in the relatively high-temperature region in Extended Data Fig. 4a are approximately proportional to the corresponding activation energies. The slope decreases with increasing Y concentration, which implies that Y-substitution increases the bulk conductivity and reduces the activation energy of carriers. More explicitly, in order to estimate and provide the activation energies of SmB₆ and Sm_{0.8}Y_{0.2}B₆, only the bulk conductance channel should be taken into account. Thus, based on a simple parallel conductance model (total $G = G_{\text{bulk}} + G_{\text{surface}}$) below the temperature at which the Kondo gap is completely open (roughly 40 K)^{14,16,21,35,41}, we plot $G - G_{\text{surface}}$ (logarithmic scale) against $1/T$ in Extended Data Fig. 4b, where G_{surface} is modelled as a linear function of temperature⁴¹, and $G - G_{\text{surface}}$ is normalized by *G* at 300 K. Now the slopes of $G - G_{\text{surface}}$ in Extended Data Fig. 4b correspond to the activation energies of pure SmB₆ and Sm_{0.8}Y_{0.2}B₆, which are found to be 3.0 meV and 2.2 meV, respectively.

Details of the point-contact spectroscopy measurements. PCAR measurements were carried out using a probe built in-house and designed for operation in a physical property measurement system (Quantum Design). Using a mechanically sharpened tip, point-contact junctions with a contact resistance of few ohms were achieved by gently approaching the tip onto the surface of the heterostructure at 2 K. In order to demonstrate the robustness of perfect Andreev reflection observed in the SmB₆ (20–30 nm)/YB₆ heterostructures, we made multiple contact measurements by lifting up the PtIr tip and repositioning it to land at other spots (position 1–3) on the same samples. As shown in Extended Data Fig. 5, in each set of such measurements, we consistently obtained conductance doubling for all contacts made on SmB₆ (20–30 nm)/YB₆ heterostructures despite the expected local variation in the surface microstructure.

Conductance enhancement against Z-barrier strength. There are many factors that cause scattering and thus contribute to the barrier strength *Z* in the standard BTK theory^{9,10}. In point-contact spectroscopy experiments, it is often difficult to avoid the formation of an oxide layer at the surface. Even when the interface is

formed in situ under vacuum for thin-film devices, the interfaces are defined as where the two disparate materials meet: the difference in the crystal structure and the atomic-level surface microstructure, including facets and terminations, can lead to structural and compositional disorder and defects serving as scattering centres. Mechanical point contacts have an added complication due to local deformation of the tip. Furthermore, Fermi velocity mismatch also affects the reflection and transmission probabilities. Therefore, Z is finite for almost all normal metal–topologically trivial superconductor junctions, which leads to conductance enhancements of considerably less than two.

To illustrate the uniqueness of the perfect Andreev reflection that is evident here in the doubled conductance (normalized $dI/dV = 2$) and the difficulty in general in observing such a high conductance enhancement, we surveyed the open literature on point-contact spectroscopy measurements on various superconductors. Because Z is a primary parameter associated with the conductance enhancement in the standard BTK theory⁹, we plot normalized dI/dV at zero bias against Z (Fig. 3). We looked at over 250 publications on point-contact spectroscopy measurements and selected data points from 44 reports using three criteria: (1) the value of Z is extracted using a BTK fit; (2) the conductance enhancement is larger than 1 (normalized $dI/dV \geq 1$), which indicates that a particular junction is not in the tunnel-dominant regime; and (3) the conductance enhancement is not governed by any zero-bias conductance peak due to a nodal order parameter. For the plot in Fig. 3 we display the data points in the range of $0 < Z \leq 0.8$. Detailed information—including the types of superconductors, contacts and their references—are summarized in Supplementary Table.

Comparison of dI/dV spectra in standard BTK and Dirac–BTK models. Extended Data Fig. 6a shows a comparison of dI/dV curves according to the standard BTK and the Dirac–BTK models for different Z values, from which it can be clearly seen how the dI/dV spectrum is modified by changing the barrier strength Z . In the standard BTK model, the conductance within the superconducting gap gradually decreases with increasing Z , whereas the dI/dV spectra in the Dirac–BTK model remain unchanged regardless of the value of Z , as theoretically described in the main text. Such dependency is also captured in the curves in Fig. 3 (that is, normalized dI/dV at zero bias against Z according to the standard BTK and the Dirac–BTK models). Extended Data Fig. 6b shows the comparison of the Dirac–BTK and the standard BTK fits to the dI/dV spectrum of a PtIr–SmB₆ (20 nm)/YB₆ contact. When the standard BTK model is used, as expected, the best fit is obtained by setting $Z = 0$, which then provides an identical fit to the Dirac–BTK (with the same Δ). If we use a more realistic value, $Z = 0.39$, the standard BTK gives a fit with considerable deviation from the experimental curve. As discussed in the main text, $Z = 0.39$ was extracted from spectra of heterostructures that are similar in terms of materials but do not have complete topological protection, namely SmB₆ (10 nm)/YB₆ and Y-substituted SmB₆/YB₆ heterostructures. The plot clearly demonstrates that the standard BTK model with a finite and realistic Z cannot reproduce the experimental data that show the perfect Andreev reflection.

Magnetic-field-dependent dI/dV spectra of a point contact with a SmB₆/YB₆ heterostructure. Applying a magnetic field can break time-reversal symmetry, and the effect can be used as a signature of the perfect Andreev reflection due to Klein tunnelling. We measured the field-dependent dI/dV spectrum of a device with a thin-film Au layer as a normal metal (that is, Au–SmB₆/YB₆ structure; see Extended Data Fig. 3) that provides a stable contact under an applied magnetic field, as opposed to a point-contact junction which can potentially suffer from magnetostriiction. As shown in Extended Data Fig. 7a, the enhancement of conductance is gradually suppressed with increasing magnetic field in both out-of-plane and in-plane field configurations, but the normalized dI/dV at zero bias decreases more quickly when the magnetic field is applied along the out-of-plane direction compared to when it is applied in-plane (Extended Data Fig. 7b). However, the decreasing trend of the superconducting gap (Δ) due to applied field is approximately the same for the out-of-plane and in-plane directions (inset of Extended Data Fig. 7b). The fact that the conductance is suppressed more quickly under the out-of-plane field thus cannot be explained solely by field-induced diminishing of superconductivity in the SmB₆/YB₆ heterostructure.

The effect of magnetic field on the helical surface states depends on factors such as the direction of the field, the position of the Fermi level relative to the Dirac point, and the magnitude of the effective g -factor. Applying the magnetic field parallel to the surface will distort and shift the Dirac cone, but without affecting the spin–momentum locking at the Fermi level^{42,43}. However, a magnetic-field component perpendicular to the surface will open a gap at the Dirac point and a back-scattering channel by inducing a z -component of the electron spins^{42–44}. In other words, we expect considerable suppression of the conductance when the field is applied out of plane, which is consistent with our observation here.

The observed suppression, however, is not especially pronounced in either direction, and we attribute this to the small effective g -factor of the surface states in SmB₆. The size of the opened gap or the shift in Fermi surface (Δ_B) due to the magnetic field B is proportional to the Zeeman energy, $\Delta_B = g_{\text{eff}}\mu_B B$, where g_{eff}

is the effective g -factor of surface states and μ_B is the Bohr magneton⁴⁵. Thus, for sufficiently small g_{eff} , the application of B does not weaken topological protection substantially, provided that the Fermi level is sufficiently far away from the Dirac point. To the best of our knowledge, the effective g -factor for the surface states of SmB₆ has not been reported, but the value for the bulk states of SmB₆ has been estimated to be around 0.1^{46,47}. It has been reported that the effective g -factor of surface states of Bi₂Se₃ is similar to the bulk value in Bi₂Se₃ ($g_{\text{eff}} \approx 50$)⁴⁸. In the absence of a directly measured value for SmB₆ and assuming that its behaviour is similar to that of Bi₂Se₃, we take the g -factor of the surface states of SmB₆ to also be around 0.1.

Recent magnetoresistance studies on SmB₆ also suggest a small effective g -factor for the surface states of SmB₆^{47,49,50}. For example, a very weak field-dependence of the resistance at low temperatures (for instance, $\Delta R/R \approx 2\%$ for 80 T at 1.39 K) has been reported⁴⁷, which suggests that the surface states of SmB₆ are extremely robust against an applied magnetic field. This is consistent with the gradual suppression of conductance enhancement by magnetic fields that we have observed here.

Conductance doubling and conductance dip near the gap. To the best of our knowledge, there have been only two reports in the literature in which the observed conductance enhancement is larger than 1.9. They are both on Nb–Cu point contacts^{28,29} (also see Fig. 3 and Supplementary Table). The spectra showing conductance doubling therein are reproduced in Extended Data Fig. 8, and one of our PtIr–SmB₆ (20 nm)/YB₆ spectra is also shown in the figure for comparison. The reported Nb–Cu spectra exhibit distinctive features—namely, conductance dips near the bias voltage corresponding to the superconducting gap energy of Nb (indicated by arrows in Extended Data Fig. 8). These dips cannot be reproduced using the standard BTK theory alone. A model has been proposed to account for the dips that are intimately tied to the conductance doubling²⁹. In this model, when Z is exceptionally small due to a negligible Fermi velocity mismatch—as in the special case of Nb–Cu junctions—the interface becomes effectively transparent, which enables the superconducting proximity effect to create a region in the normal metal side with a superconducting order parameter (Δ_{prox}) that is smaller than the order parameter of the superconductor. In such an instance, the Andreev reflection process is limited to the energy of incident particles within $|\Delta_{\text{prox}}|$. According to the model put forth in ref. 29, because the quasiparticles in the proximitized layer on the normal metal side can enter the superconductor side only when their energy is outside the energy gap of the superconductor, the dI/dV spectrum develops large conductance dips near voltages that roughly correspond to the gap energy of the superconductor. Therefore, we attribute the substantial dip feature to $Z \approx 0$ in the case of Nb–Cu junctions. The absence of such a feature in our results thus indicates that the perfect conductance doubling observed in the PtIr–SmB₆ (20–30 nm)/YB₆ junctions is of a different origin compared to that in the Nb–Cu contacts. In the case of a contact between PtIr and SmB₆, a substantial barrier is expected just on the basis of the substantial Fermi velocity mismatch between them (the Fermi velocity of the surface states of SmB₆ is $< 10^5$ m s^{−1})^{17,18,21}. This underscores the need for an alternative model to explain the perfect Andreev reflection observed in the PtIr–SmB₆/YB₆ heterostructures here.

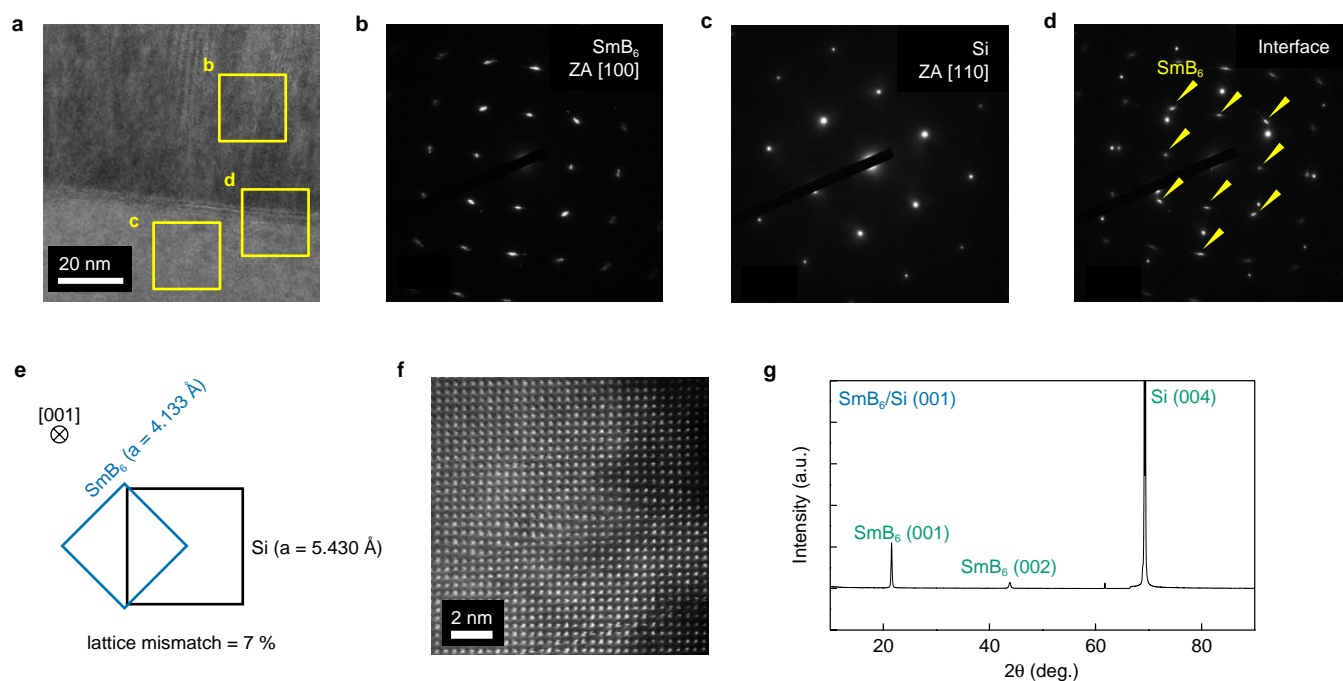
Note that shallow dips observed in the dI/dV spectra of our PtIr–SmB₆ (20 nm)/YB₆ junctions are common to dI/dV spectra of various normal metal–superconductor junctions (for example, refs 10,51–54). They are attributed to the inhomogeneous nature of point contact, which can consist of many parallel channels—in some of which excessive current can flow^{10,55}.

Data availability

The data that support the findings of this study are available within the paper. Additional data are available from the corresponding authors upon reasonable request.

- Yong, J. et al. Robust topological surface state in Kondo insulator SmB₆ thin films. *Appl. Phys. Lett.* **105**, 222403 (2014).
- Li, Y., Ma, Q., Huang, S. X. & Chien, C. L. Thin films of topological Kondo insulator candidate SmB₆: strong spin–orbit torque without exclusive surface conduction. *Sci. Adv.* **4**, eaap8294 (2018).
- Ohring, M. *Materials science of thin films* 2nd edn (Academic, 2001).
- Schneider, R., Geerk, J. & Rietschel, H. Electron tunnelling into a superconducting cluster compound: YB₆. *Eur. Phys. J. Lett.* **4**, 845–849 (1987).
- Sluchanko, N. et al. Lattice instability and enhancement of superconductivity in YB₆. *Phys. Rev. B* **96**, 144501 (2017).
- Dynes, R. C., Narayana, V. & Garno, J. P. Direct measurement of quasiparticle-lifetime broadening in a strong-coupled superconductor. *Phys. Rev. Lett.* **41**, 1509–1512 (1978).
- Mazin, I. I., Golubov, A. A. & Nadgorny, B. Probing spin polarization with Andreev reflection: a theoretical basis. *J. Appl. Phys.* **89**, 7576–7578 (2001).
- Wolgast, S. et al. Low-temperature surface conduction in the Kondo insulator SmB₆. *Phys. Rev. B* **88**, 180405 (2013).
- Taskin, A. A. et al. Planar Hall effect from the surface of topological insulators. *Nat. Commun.* **8**, 1340 (2017).
- Wang, L.-X. et al. Zeeman effect on surface electron transport in topological insulator Bi₂Se₃ nanoribbons. *Nanoscale* **7**, 16687–16694 (2015).

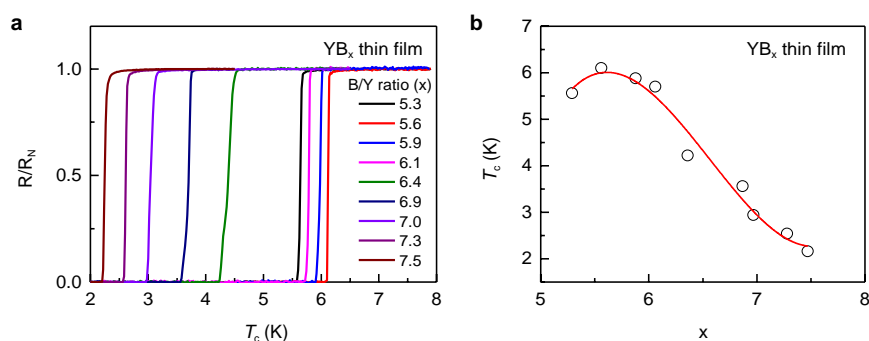
44. Chang, C.-Z., Wei, P. & Moodera, J. S. Breaking time reversal symmetry in topological insulators. *MRS Bull.* **39**, 867–872 (2014).
45. Fu, Y.-S. et al. Observation of Zeeman effect in topological surface state with distinct material dependence. *Nat. Commun.* **7**, 10829 (2016).
46. Erten, O., Ghaemi, P. & Coleman, P. Kondo breakdown and quantum oscillations in SmB_6 . *Phys. Rev. Lett.* **116**, 046403 (2016).
47. Wolgast, S. et al. Reduction of the low-temperature bulk gap in samarium hexaboride under high magnetic fields. *Phys. Rev. B* **95**, 245112 (2017).
48. Analytis, J. G. et al. Transport in the quantum limit by two-dimensional Dirac fermions in a topological insulator. *Nat. Phys.* **6**, 960–964 (2010).
49. Thomas, S. et al. Weak antilocalization and linear magnetoresistance in the surface state of SmB_6 . *Phys. Rev. B* **94**, 205114 (2016).
50. Biswas, S. et al. Robust local and nonlocal transport in the topological Kondo insulator SmB_6 in the presence of a high magnetic field. *Phys. Rev. B* **92**, 085103 (2015).
51. Gonnelli, R. S. et al. Temperature and junction-type dependency of Andreev reflection in MgB_2 . *J. Phys. Chem. Solids* **63**, 2319–2323 (2002).
52. Li, Z.-Z. et al. Andreev reflection spectroscopy evidence for multiple gaps in MgB_2 . *Phys. Rev. B* **66**, 064513 (2002).
53. Park, W. K., Greene, L. H., Sarrao, J. L. & Thompson, J. D. Andreev reflection at the normal-metal/heavy-fermion superconductor CeCoIn_5 interface. *Phys. Rev. B* **72**, 052509 (2005).
54. Zhang, X. et al. Evidence of a universal and isotropic $2\Delta/k_B T_C$ ratio in 122-type iron pnictide superconductors over a wide doping range. *Phys. Rev. B* **82**, 020515 (2010).
55. Sheet, G., Mukhopadhyay, S. & Raychaudhuri, P. Role of critical current on the point-contact Andreev reflection spectra between a normal metal and a superconductor. *Phys. Rev. B* **69**, 134507 (2004).



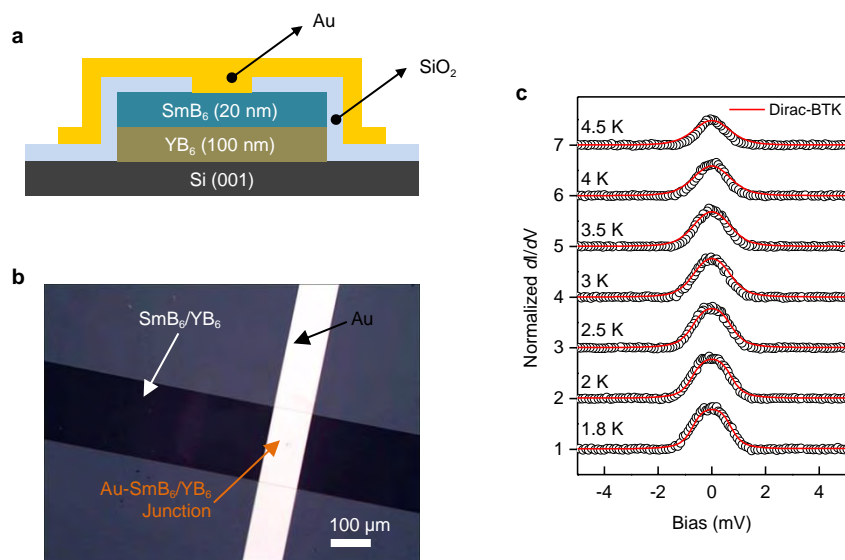
Extended Data Fig. 1 | Structural characterization of SmB₆ thin films.

a, High-resolution cross-sectional transmission electron microscopy image of a SmB₆ thin film. The yellow squares correspond to the regions of the SAED measurements shown in **b–d**. **b–d**, SAED measurements of SmB₆ (**b**), Si substrate (**c**) and SmB₆/Si interface regions (**d**). ZA, zone

axis. **e**, Epitaxial relationship between the SmB₆ and the Si substrate. **f**, Aberration-corrected scanning transmission electron microscopy cross-sectional image of a SmB₆ thin film. **g**, $\theta-2\theta$ X-ray diffraction pattern of a SmB₆ thin film on a Si (001) substrate.

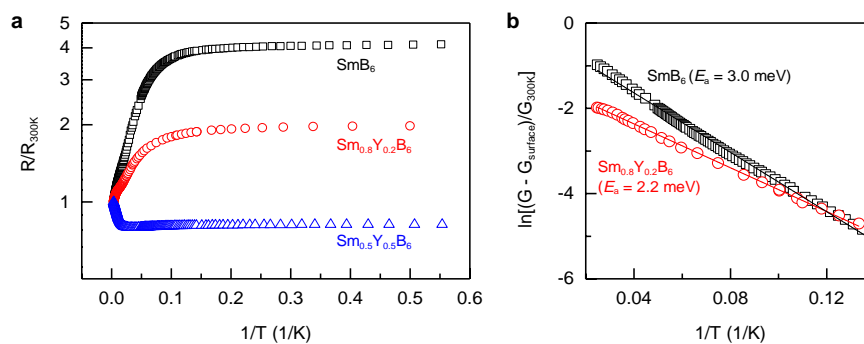


Extended Data Fig. 2 | Superconducting transition temperature (T_c) of YB_x thin films. a, Temperature-dependent resistance curves of YB_x thin films with different stoichiometric B/Y ratios. **b,** Change in T_c as a function of stoichiometric B/Y ratio (x).



Extended Data Fig. 3 | Au-SmB₆/YB₆ thin film junctions. **a**, Cross-sectional schematic of a Au-SmB₆ (20 nm)/YB₆ (100 nm) structure. **b**, Optical microscopy image of the device. **c**, Normalized dI/dV spectra of the Au-SmB₆/YB₆ structure at different temperatures. The red lines are

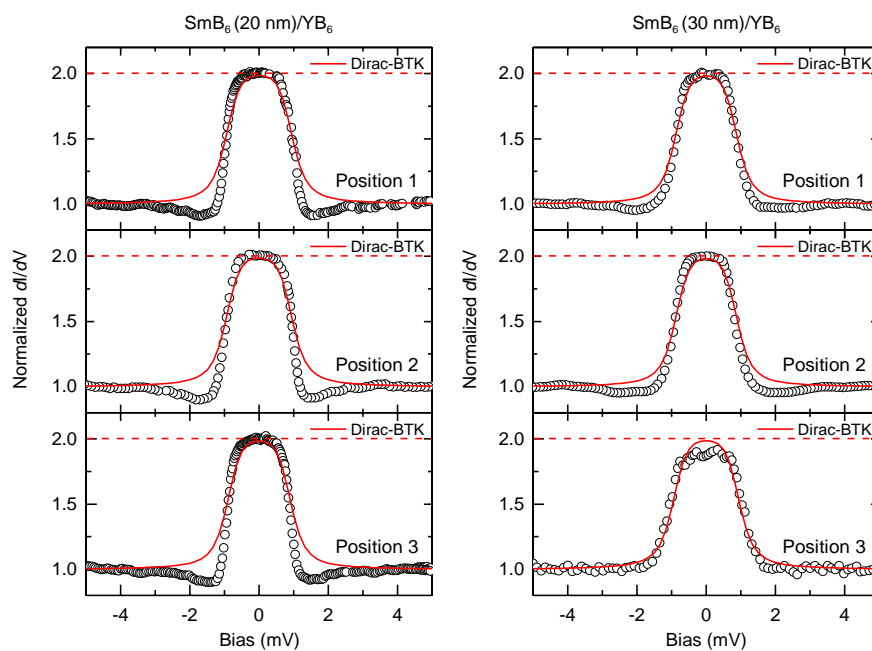
fits using the Dirac-BTK model. The normalized dI/dV curves at 1.8 K are plotted using the obtained values, whereas the other curves are vertically shifted for clarity.



Extended Data Fig. 4 | Yttrium-substituted SmB_6 thin films.

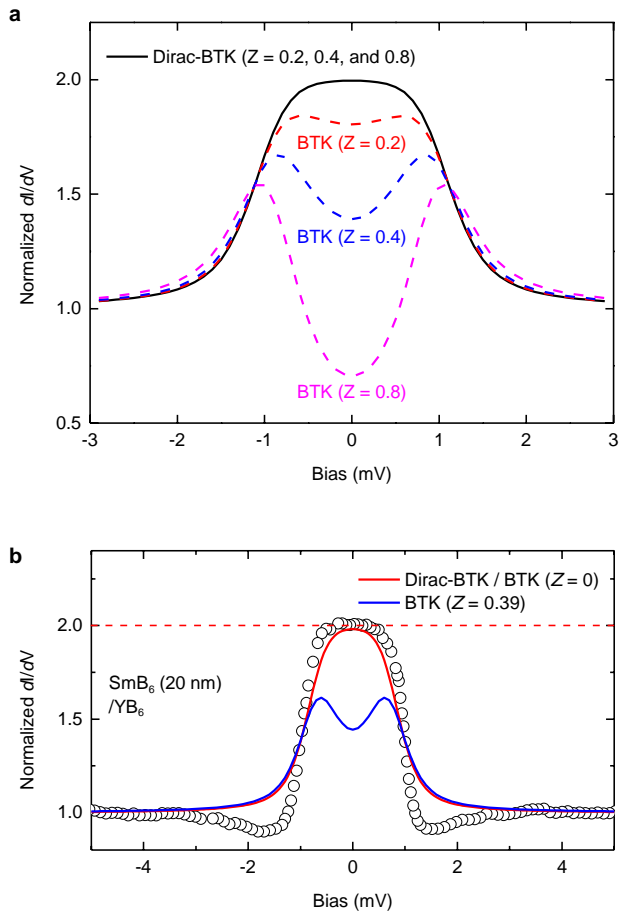
a, Comparison of $\log R$ against $1/T$ plots of SmB_6 , and 20% and 50% Y-substituted SmB_6 (that is, $Sm_{0.8}Y_{0.2}B_6$ and $Sm_{0.5}Y_{0.5}B_6$, respectively). The resistance values are normalized by their values at 300 K. The positive linear slopes at in the relatively high-temperature regions are roughly

proportional to the activation energy. **b**, $G - G_{surface}$ (logarithmic scale, normalized by the conductance at 300 K) plotted against $1/T$ for pure SmB_6 (black squares) and $Sm_{0.8}Y_{0.2}B_6$ (red circles). The slopes of the linear fits (black and red lines) correspond to the activation energies (E_a) of pure SmB_6 and $Sm_{0.8}Y_{0.2}B_6$, and are 3.0 meV and 2.2 meV, respectively.

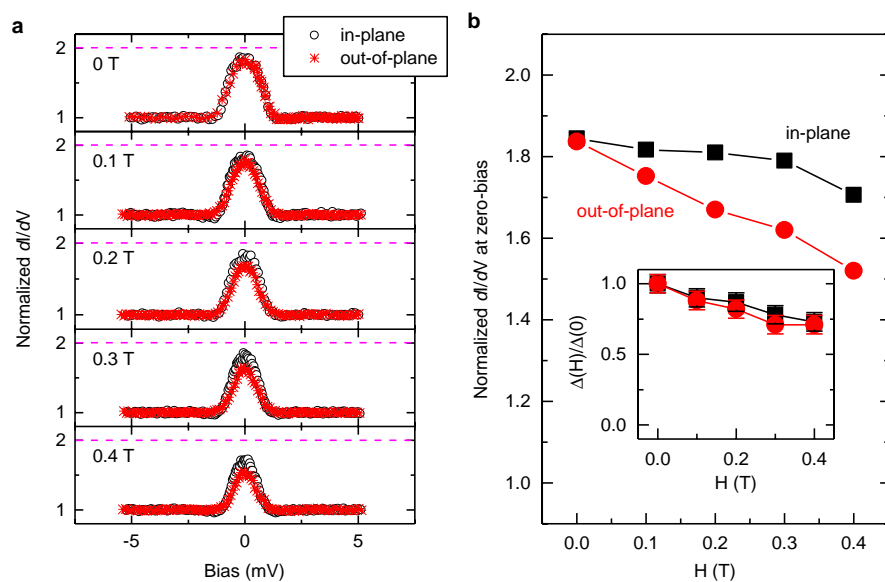


Extended Data Fig. 5 | Robustness of perfect Andreev reflection.
Point-contact spectra obtained at different positions (1, 2 and 3, which are roughly 1 mm apart from each other) on SmB_6/YB_6 heterostructures

with 20-nm-thick SmB_6 (left) and 30-nm-thick SmB_6 (right). Conductance doubling is consistently observed at all positions in the dI/dV spectra of the SmB_6/YB_6 heterostructures.

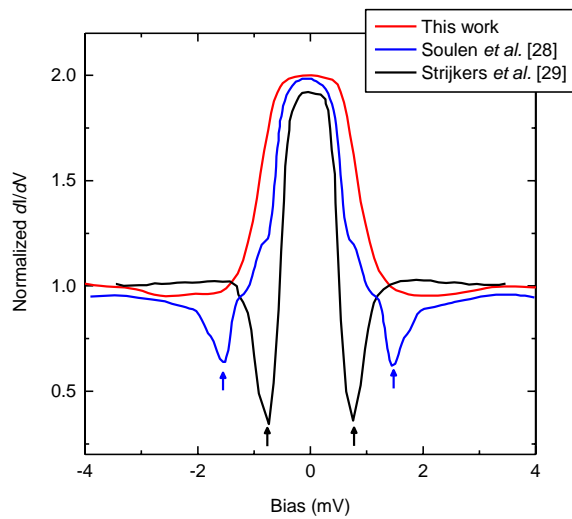


Extended Data Fig. 6 | Standard BTK compared with Dirac-BTK models. **a**, Comparison of calculated dI/dV spectra with the standard BTK and the Dirac-BTK models for $Z = 0.2, 0.4$ and 0.8 ($\Delta = 1 \text{ meV}$). **b**, Comparison of the Dirac-BTK and the standard BTK fits to the experimental dI/dV spectrum of a PtIr-SmB₆ (20 nm)/YB₆ contact (Fig. 1c). The red curve is the theoretical conductance curve in the Dirac-BTK model and the standard BTK model with $Z = 0$. Both appear identical, as expected, for the same Δ (here 0.77). The blue curve is the theoretical standard BTK curve with $\Delta = 0.77$ and $Z = 0.39$, this Z value is assessed from contacts to other heterostructures in this study that do not exhibit perfect Andreev reflection (that is, those with thin SmB₆ (10 nm) and Y-substituted SmB₆). The effect of nullifying Z by incorporation of a Dirac material in the Andreev reflection process is clearly seen.



Extended Data Fig. 7 | Magnetic-field-dependent dI/dV spectra. **a**, dI/dV spectra of Au-SmB₆/YB₆ device under a magnetic field applied along the in-plane and out-of-plane directions. **b**, Normalized dI/dV at zero bias as a function of magnetic field. The inset shows superconducting order

parameter (Δ) as a function of magnetic field normalized by Δ at 0 T ($\Delta(0)$). Δ was estimated as the bias voltage point at which the maximum first derivative of each dI/dV spectrum occurs under different magnetic fields.



Extended Data Fig. 8 | Experimentally observed conductance doubling.

Comparison of the normalized dI/dV spectrum obtained from the PtIr-SmB₆ (20 nm)/YB₆ junction in this work (red line, experimental data) with the reported point-contact spectra obtained from Nb-Cu junctions^{28,29}. The arrows indicate conductance dips near the Δ . Such dips are not present in our spectrum.

Enhanced intrinsic photovoltaic effect in tungsten disulfide nanotubes

Y. J. Zhang^{1,2*}, T. Ideue³, M. Onga³, F. Qin³, R. Suzuki³, A. Zak⁴, R. Tenne⁵, J. H. Smet² & Y. Iwasa^{3,6}

The photovoltaic effect in traditional *p–n* junctions—where a *p*-type material (with an excess of holes) abuts an *n*-type material (with an excess of electrons)—involves the light-induced creation of electron–hole pairs and their subsequent separation, generating a current. This photovoltaic effect is particularly important for environmentally benign energy harvesting, and its efficiency has been increased dramatically, almost reaching the theoretical limit¹. Further progress is anticipated by making use of the bulk photovoltaic effect (BPVE)², which does not require a junction and occurs only in crystals with broken inversion symmetry³. However, the practical implementation of the BPVE is hampered by its low efficiency in existing materials^{4–10}. Semiconductors with reduced dimensionality² or a smaller bandgap^{4,5} have been suggested to be more efficient. Transition-metal dichalcogenides (TMDs) are exemplary small-bandgap, two-dimensional semiconductors^{11,12} in which various effects have been observed by breaking the inversion

symmetry inherent in their bulk crystals^{13–15}, but the BPVE has not been investigated. Here we report the discovery of the BPVE in devices based on tungsten disulfide, a member of the TMD family. We find that systematically reducing the crystal symmetry beyond mere broken inversion symmetry—moving from a two-dimensional monolayer to a nanotube with polar properties—greatly enhances the BPVE. The photocurrent density thus generated is orders of magnitude larger than that of other BPVE materials. Our findings highlight not only the potential of TMD-based nanomaterials, but also more generally the importance of crystal symmetry reduction in enhancing the efficiency of converting solar to electric power.

Tungsten disulfide (WS₂) and other group-VI-B TMDs are two-dimensional semiconductors with relatively small bandgaps of 1.2–2.1 eV (refs ^{11,12}). In their most stable 2H phase, the unit cell of bulk TMDs is centrosymmetric and the material belongs to the *D*_{6h} point group. However, because the unit cell is composed of a bilayer, it is

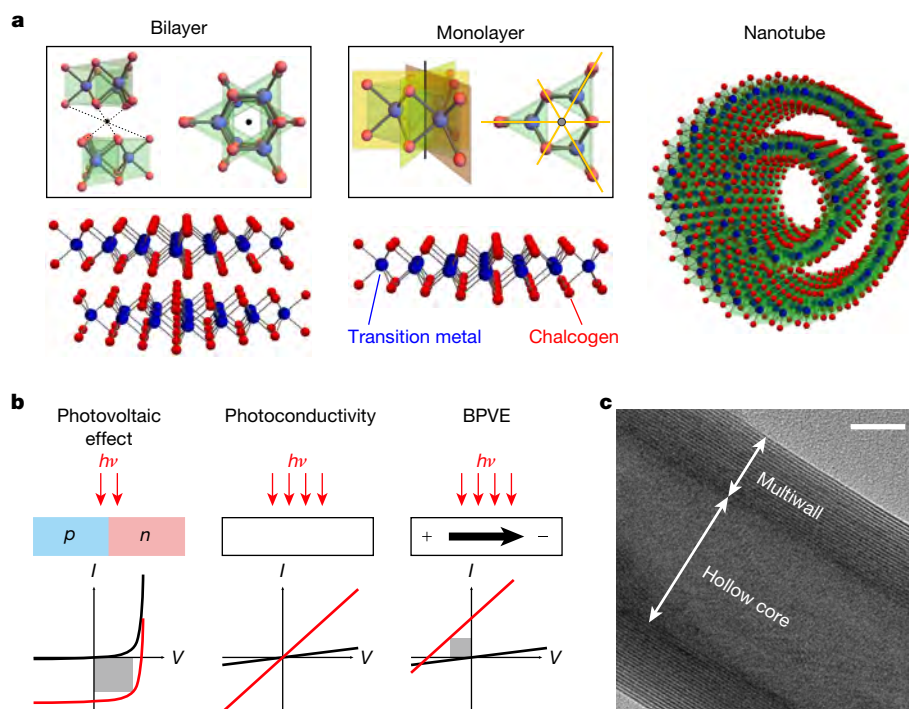


Fig. 1 | Photovoltaic response in TMD nanomaterials. **a**, Illustrations of the crystal structure of WS₂ for devices of different symmetry: a bilayer, a monolayer and a multiwall nanotube. The black dots in the bilayer mark the centre of the spatial inversion symmetry. The yellow planes and the black pole at the cross-section of the yellow planes in the monolayer indicate the mirror planes and the threefold rotation axis, respectively. **b**, Comparison of three different light-to-current conversion mechanisms. Black and red lines illustrate typical *I–V* characteristics

in the dark and under illumination, respectively. The grey rectangles represent the generated electrical power. Generally, the photovoltaic effect requires a junction between a *p*-type and an *n*-type material; if there is no junction, only a change in conductivity occurs (without electrical power generation). For materials without inversion symmetry, the BPVE can occur and generate electrical power without a junction. **c**, TEM image of a hollow-core multiwall WS₂ nanotube. The white scale bar represents 10 nm.

¹The Institute of Scientific and Industrial Research, Osaka University, Osaka, Japan. ²Max Planck Institute for Solid State Research, Stuttgart, Germany. ³Department of Applied Physics and Quantum-Phase Electronics Center (QPEC), The University of Tokyo, Tokyo, Japan. ⁴Faculty of Sciences, HIT-Holon Institute of Technology, Holon, Israel. ⁵Department of Materials and Interfaces, Weizmann Institute of Science, Rehovot, Israel. ⁶RIKEN Center for Emergent Matter Science (CEMS), Wako, Japan. *e-mail: y.zhang@ikf.mpg.de

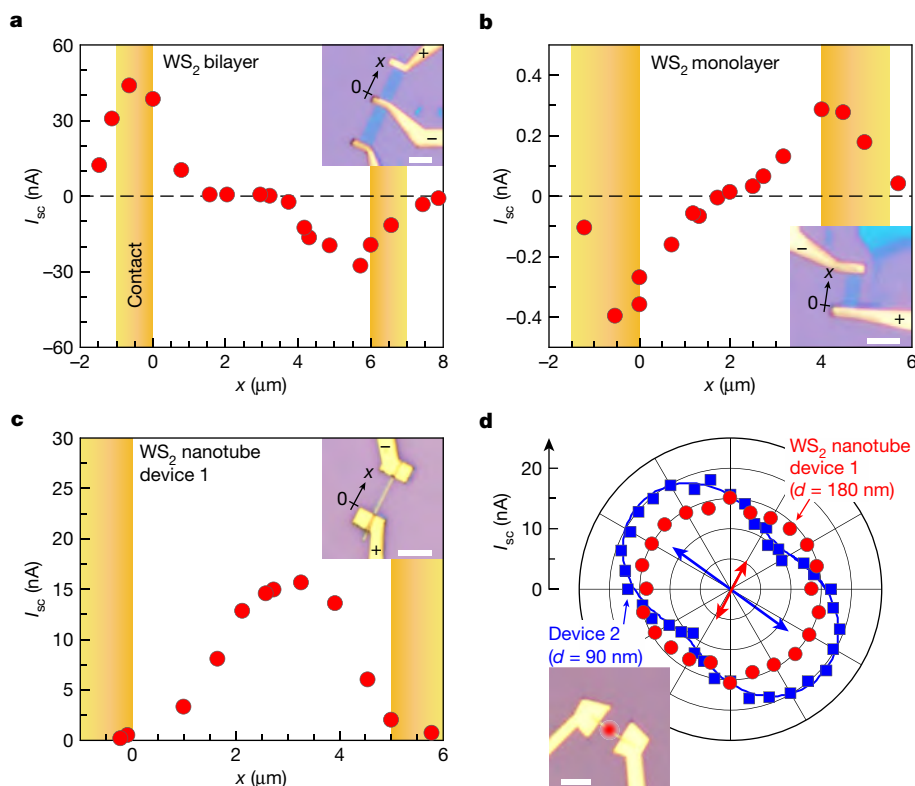


Fig. 2 | The photovoltaic response obtained with WS₂-based devices of different crystal symmetry. White scale bars in the optical micrographs of the devices (insets) represent 4 μm . The excitation laser wavelength was 632.8 nm in all cases. I_{sc} is positive when the current runs from the + electrode to the – electrode shown in the insets. **a**, The dependence of I_{sc} on the position of the laser spot in a WS₂ bilayer device. x is the distance between the laser spot and one of the electrodes (see inset). When the laser spot illuminates the contact area, the conventional Schottky barrier photovoltaic effect and/or the photothermal effect is observed. **b**, The dependence of I_{sc} on the position of the laser spot in a WS₂ monolayer

device. Here, too, laser illumination triggers a photovoltaic response only near the contact area. **c**, The dependence of I_{sc} on the position of the laser spot in a WS₂ nanotube device. The main response occurs when the laser spot illuminates the centre of the device away from the contacts. Hence, this is a bulk photovoltaic effect. **d**, Polar plot of I_{sc} in WS₂ nanotube devices. The blue line is a $\cos(2\theta)$ fit, where θ is the linear polarization angle of the laser. The red circle in the optical micrograph (inset) marks the position of the laser spot. The arrows in the polar diagram represent the direction of the tube axis for each nanotube device. d is the (outer) diameter of each nanotube.

possible to break inversion symmetry simply by isolating an individual layer (Fig. 1a). Monolayers belong to the non-centrosymmetric and non-polar D_{3h} point group. In such monolayers, the BPVE—which converts solar power into electric power without the need for a p – n junction (Fig. 1b)—can be present provided that the incident light is linearly polarized along specific directions, whereas circularly polarized light is not allowed to do so³. When threefold rotational and mirror symmetries of the 2D sheets are removed by curvature or strain, light can induce the BPVE irrespective of its polarization characteristics⁵. Here we investigate the BPVE in WS₂ devices with successively lower crystal symmetry, namely centrosymmetric bilayers, non-centrosymmetric non-polar monolayers and, finally, non-centrosymmetric polar nanotubes^{16–20}. The latter are multiwall nanotubes with a hollow core, as illustrated in a representative transmission electron microscopy (TEM) image (Fig. 1c). The devices have chromium and gold electrodes, and we refer to the Methods and Extended Data Fig. 1 for characterization of the nanotubes.

The short-circuit current (I_{sc}) under laser illumination is an important figure-of-merit for evaluating the photovoltaic effect. We have measured this current for our WS₂ devices with different crystal symmetry (Fig. 2a–c). For each device, we scanned the laser spot from one electrode to the other in order to distinguish the BPVE from the Schottky barrier photovoltaic effect and from the photothermal effect near contacts^{21,22}. In the WS₂ bilayer device, a photovoltaic response occurs only when the contacts are illuminated by laser light, whereas no signal appears when just the WS₂ flake itself is illuminated (Fig. 2a). This is consistent with the symmetry requirements for observing the BPVE. The monolayer device behaves similarly to the bilayer device,

showing no notable photovoltaic response away from the contacts (Fig. 2b). I_{sc} also remains small when the laser spot is fixed near the centre of the WS₂ flake while rotating the linear polarization direction. There are no symmetry-related arguments that prevent the BPVE in a monolayer device under linearly polarized light, yet the experiment clearly indicates that, if present at all, its amplitude is very small and buried within the photovoltaic response that results from the Schottky barriers and/or the photothermal effect.

In sharp contrast, I_{sc} in the WS₂ nanotube device increases substantially when the middle of the nanotube is illuminated (Fig. 2c). Similar data for a second WS₂ nanotube device—as well as a summary of the variation in the BPVE's amplitude in five different devices—can be found in Extended Data Fig. 2 and Extended Data Table 1, respectively. The large increase in I_{sc} away from the contacts in nanotube devices cannot be explained simply by a difference in the amount of absorbed light. Even though the nanotubes may indeed absorb more light than would a monolayer owing to their multiwall nature, when one considers the nanotube diameter versus the laser spot size, the absorbed light intensity is at best twice as large as that in a WS₂ monolayer (see Methods). However, the observed difference in I_{sc} between the WS₂ nanotube devices (roughly 15 nA; Fig. 2c) and the WS₂ monolayer devices (less than 0.1 nA; Fig. 2b) amounts to several orders of magnitude. The reduction in the crystal symmetry beyond mere non-centrosymmetry is apparently crucial.

The lower crystal symmetry of nanotubes also manifests itself in a study of the dependence of I_{sc} on the linear polarization direction of the incident laser beam. For WS₂ monolayers obeying the symmetries of the D_{3h} point group, the BPVE should drop to zero for specific

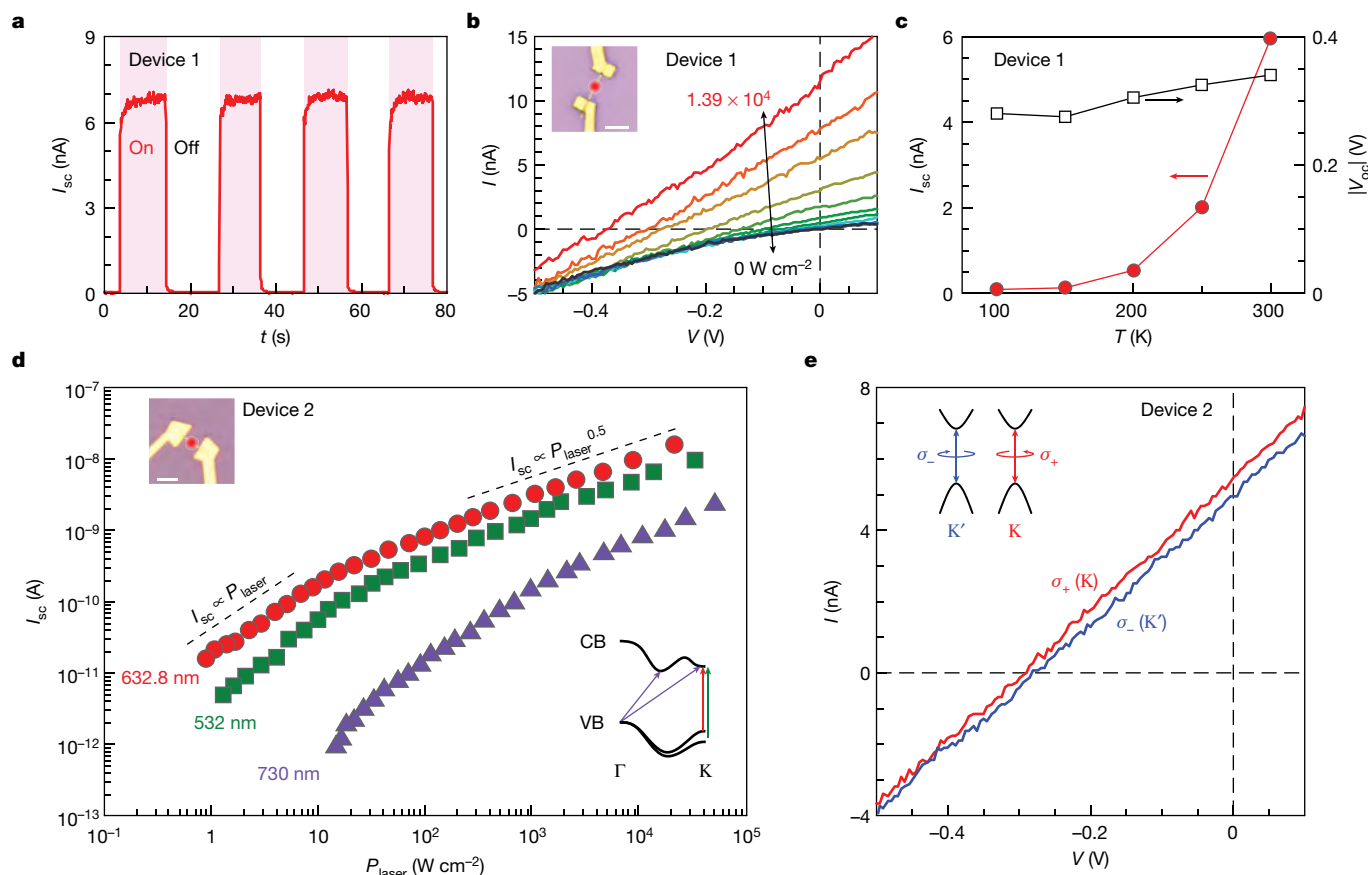


Fig. 3 | Photovoltaic response in WS₂ nanotubes. **a**, I_{sc} record during on/off cycles of an incident laser with a wavelength of 632.8 nm and a laser power of $1.39 \times 10^4 \text{ W cm}^{-2}$. **b**, I - V characteristics recorded at different illumination intensities. The laser wavelength was 632.8 nm. Inset, optical micrograph of the device. The laser spot was fixed to the centre of the nanotube. The white scale bar represents $4 \mu\text{m}$. **c**, I_{sc} (filled circles) and V_{oc} (open squares) in WS₂ nanotube device 1, recorded at different temperatures. The incident laser wavelength and the power were 632.8 nm and $1.39 \times 10^4 \text{ W cm}^{-2}$, respectively. **d**, Dependence of I_{sc} on P_{laser} for three different wavelengths. The black dashed lines are guides to

the eye. The bottom right inset illustrates possible excitation paths from the valence band (VB) to the conduction band (CB) for each wavelength, using the same colour scheme as the data points. Γ and K represent the centre and the edge, respectively, of the hexagonal Brillouin zone. Inset, optical micrograph of the device. The white scale bar represents $4 \mu\text{m}$. **e**, I - V curves for laser light with right-handed (σ_+) and left-handed (σ_-) circular polarization. The incident laser wavelength and the power were 632.8 nm and $1.39 \times 10^4 \text{ W cm}^{-2}$, respectively. The inset depicts the valley-contrasting optical selection rules. Light with σ_+ (or σ_-) circular polarization is required to excite carriers in the K (or K') valley.

polarization directions of the incident light³. However, experimentally we are unable to resolve this dependence on the polarization direction, probably because the BPVE is too small to be seen in monolayers. By contrast, mirror and rotational symmetries are broken in nanotubes because of the curvature, and hence the BPVE should persist for all linear polarization directions³. This behaviour is clearly demonstrated in Fig. 2d. This figure shows I_{sc} in a polar diagram as a function of the linear polarization angle, θ . In device 1, I_{sc} is almost constant with angle. In device 2, I_{sc} remains large for all angles, but some angular dependence is apparent. The discrepancy between the two devices can probably be attributed to differences in the exact symmetry of the nanotubes (Methods). The anisotropy of the absorption coefficient for devices with a one-dimensional geometry may also play a part. This anisotropy is less pronounced in device 1, because its diameter is twice as large as that of device 2 (180 nm versus 90 nm).

The BPVE in WS₂ nanotubes is fairly robust both qualitatively and quantitatively. Figure 3a illustrates the change in I_{sc} upon multiple on-off cycles of the laser; I_{sc} returns to a similar value whenever laser illumination is on. Figure 3b displays the I - V characteristics for a range of laser power levels (P_{laser}). Both I_{sc} and the open-circuit voltage (V_{oc}) change monotonically with P_{laser} . They are also affected by the temperature (Fig. 3c). The large drop in I_{sc} with reduced temperature cannot be explained simply by a reduction in the absorption coefficient owing to the blue shift of the band gap as temperature decreases (Methods).

Figure 3d summarizes the dependence of I_{sc} on P_{laser} for three different laser wavelengths in a log-log plot. At the same P_{laser} , the following inequalities hold: I_{sc} (632.8 nm) is greater than I_{sc} (532 nm), which is greater than I_{sc} (730 nm). This can be explained by considering the wavelength dependence of the absorption coefficient²⁰. Light with a wavelength of 632.8 nm (1.96 eV) is nearly resonant with the A-exciton (a specific bound state of an electron and a hole) of WS₂, and therefore produces the strongest signal. The external quantum efficiency—as deduced from the slope of a linear fit to the data—reaches a value as high as 1.3% for this wavelength in the power regime 10 W cm^{-2} . At a wavelength of 730 nm (1.70 eV), there is still a small (but unmistakable) photovoltaic response. Photons can now excite carriers only across the indirect gap (1.45 eV, or 855 nm in wavelength), as seen in the band-structure diagram in Fig. 3d. This suggests that approximately 60% of the solar electromagnetic spectrum (photon energies above 1.45 eV) can be exploited to convert the incident energy into electricity.

The power dependence in Fig. 3d reveals a crossover from a linear to a square-root dependence near $P_{laser} = 10 \text{ W cm}^{-2}$. This feature is useful in contemplating the potential mechanisms underlying the BPVE in WS₂ nanotube devices. Before addressing this, however, we note for the sake of completeness that such a dependence would not be expected for the Schottky barrier photovoltaic effect at the interface between WS₂ nanotubes and a metallic contact. This effect has been shown¹⁸ to remain linear even up to $5 \times 10^3 \text{ W cm}^{-2}$.

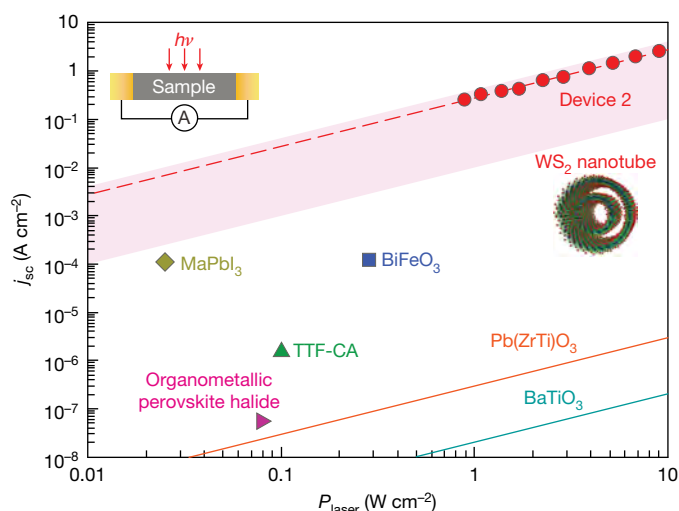


Fig. 4 | Overview of the bulk photovoltaic effect in various materials.

Data for other materials are taken from the literature (TTF-CA, ref. ⁵; BaTiO₃ and Pb(ZrTi)O₃, ref. ⁶; BiFeO₃, ref. ⁷; MaPbI₃, ref. ⁸; organometallic perovskite halide, ref. ⁹). The selected data points for these materials are for the lateral configuration (see top left inset). In the WS₂ nanotube devices, the current runs parallel to the tube axis. The red dashed line is a linear fit to the experimental data from our device 2. Red shading demarcates the variation among devices, as estimated from the data in Extended Data Table 1. The solid lines for Pb(ZrTi)O₃ and BaTiO₃ are drawn from the values in ref. ⁶.

A finite I_{sc} generated when the incident light is away from an interface could be induced by various mechanisms. In 2D TMDs, oblique incidence may produce the photogalvanic and photon drag effects^{21,23}. However, although these exist in WS₂ nanotubes as well, they play only a minor part (see Methods and Extended Data Fig. 3). Alternatively, if photocarrier generation were inhomogeneous, a net Dember effect might appear²⁴. However, despite some inhomogeneity (Extended Data Fig. 4), the results obtained with WS₂ nanotubes neither qualitatively nor quantitatively fit this mechanism (see Methods). If the inhomogeneity is accompanied by strain variations, the so-called flexophotovoltaic effect may arise and induce a BPVE²⁵. However, the crossover from a linear to a square-root dependence (Fig. 3d) points to a different origin for the BPVE. Among the many physical mechanisms that have been proposed²⁶, the ‘shift current’ model may be applicable to WS₂ nanotubes, because this model indeed does predict a power-dependence crossover as a general feature caused by a saturation of carrier excitations^{27,28}. Shift currents originate from the Berry connection of the Bloch functions when a non-zero electronic spontaneous polarization is available^{29,30}. Given that the electronic spontaneous polarization is non-zero in heteropolar nanotubes³¹, the shift current may indeed be active in WS₂ nanotubes and could explain the observed BPVE.

A more specific, qualitative shift current model has been proposed³² for tubular structures made from a heteropolar material; in this model, shift current is generated by carriers that occupy the K and K’ valleys, through momentum quantization around the circumference of the tube. Because these valleys can be addressed selectively by using circularly polarized light¹³, the shift current should change upon reversing the direction of circular polarization. For our WS₂ nanotubes, the I – V characteristics produced in response to circularly polarized light of opposite helicities (marked as σ_+ and σ_-) are plotted in Fig. 3e (see Methods for the influence of anisotropic absorption). The signs of I_{sc} and V_{oc} are the same for both helicities. Hence, the two valleys in our study apparently contribute in the same way, rather than with opposite signs. This contradicts the valley-contrasting shift current model³². The data shown in Fig. 2 indeed suggest that the symmetry reduction caused by the curved nature of nanotubes is a key factor. However, the shift current model considered only the consequences of momentum quantization in their Hamiltonian³², which is not sufficient

because curvature can induce many other effects, including effective magnetic fields, interband spin-lattice coupling, and a change in the orbital composition of the Bloch band³³. Hence, further theoretical considerations, including other curvature-related terms, are needed to clarify the specific origin of the BPVE in WS₂ nanotubes.

Finally, we compare the strength of the BPVE in our nanotubes with that reported for ferroelectric bulk materials that also exhibit a BPVE^{5–9}. For each case, the short-circuit current density (j_{sc}) is plotted against the incident P_{laser} (Fig. 4). To calculate the j_{sc} in WS₂ nanotubes, we assume a solid cylindrical cross-section instead of a hollow tube. Hence, the plotted current density represents a lower bound. Red circles are data points obtained for device 2, and red shading demarcates the variation among devices, as estimated from the data in Extended Data Table 1. We have included only data acquired on lateral geometries—such as that shown in the inset of Fig. 4—because the device configuration strongly affects the device performance (see Methods and Extended Data Fig. 5). From Fig. 4, it becomes clear that the BPVE is orders of magnitude larger in WS₂ nanotubes than in other, bulk materials. We note that the diameter of the tubes (roughly 10² nm) is much smaller than that of the laser spot (1–2 μ m); therefore, the nanotubes absorb at best about 10% of the total incident laser power. However, in Fig. 4 the total laser power per unit area of the entire laser spot is used for the abscissa, and hence the actual deviation in the efficiency of WS₂ nanotubes is one order of magnitude larger than the figure suggests.

In summary, we have observed a large BPVE in WS₂ nanotubes. To our knowledge, this is the first observation of such an effect outside the realm of bulk materials, and it demonstrates the potential of TMD nanotubes for harvesting solar energy. Moreover, the large difference in the strength of the effect (by several orders of magnitude) between the monolayer (a non-polar non-centrosymmetric structure) and the nanotube (a polar non-centrosymmetric structure) suggests that symmetry reduction and perhaps also a polar crystal structure are crucial for enhancing the BPVE.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-019-1303-3>.

Received: 22 April 2018; Accepted: 23 April 2019;

Published online 19 June 2019.

- Shockley, W. The theory of p – n junctions in semiconductors and p – n junction transistors. *ATT Tech. J.* **28**, 435–489 (1949).
- Cook, A. M., Fregoso, B. M., de Juan, F., Coh, S. & Moore, J. E. Design principles for shift current photovoltaics. *Nat. Commun.* **8**, 14176 (2017).
- Sturman, B. I. & Fridkin, V. M. *The Photovoltaic and Photorefractive Effects in Noncentrosymmetric Materials* (Gordon and Breach Science Publishers, 1992).
- Grinberg, I. et al. Perovskite oxides for visible-light-absorbing ferroelectric and photovoltaic materials. *Nature* **503**, 509–512 (2013).
- Nakamura, M. et al. Shift current photovoltaic effect in a ferroelectric charge-transfer complex. *Nat. Commun.* **8**, 281 (2017).
- Brody, P. S. High-voltage photovoltaic effect in barium-titanate and lead titanate lead zirconate ceramics. *J. Solid State Chem.* **12**, 193–200 (1975).
- Yang, S. Y. et al. Above-bandgap voltages from ferroelectric photovoltaic devices. *Nat. Nanotechnol.* **5**, 143–147 (2010).
- Xiao, Z. G. et al. Giant switchable photovoltaic effect in organometal trihalide perovskite devices. *Nat. Mater.* **14**, 193–198 (2015).
- Sun, Z. H. et al. A photoferroelectric perovskite-type organometallic halide with exceptional anisotropy of bulk photovoltaic effects. *Angew. Chem. Int. Edn* **55**, 6545–6550 (2016).
- Nechache, R. et al. Photovoltaic properties of Bi₂FeCrO₆ epitaxial thin films. *Appl. Phys. Lett.* **98**, 202902 (2011).
- Wilson, J. A. & Yoffe, A. D. Transition metal dichalcogenides discussion and interpretation of observed optical, electrical and structural properties. *Adv. Phys.* **18**, 193–335 (1969).
- Zeng, H. L. et al. Optical signature of symmetry variations and spin-valley coupling in atomically thin tungsten dichalcogenides. *Sci. Rep.* **3**, 1608 (2013).
- Xiao, D., Liu, G. B., Feng, W. X., Xu, X. D. & Yao, W. Coupled spin and valley physics in monolayers of MoS₂ and other group-VI dichalcogenides. *Phys. Rev. Lett.* **108**, 196802 (2012).
- Zhang, Y. J., Oka, T., Suzuki, R., Ye, J. T. & Iwasa, Y. Electrically switchable chiral light-emitting transistor. *Science* **344**, 725–728 (2014).
- Wu, W. Z. et al. Piezoelectricity of single-atomic-layer MoS₂ for energy conversion and piezotronics. *Nature* **514**, 470–474 (2014).

16. Tenne, R., Margulis, L., Genut, M. & Hodes, G. Polyhedral and cylindrical structures of tungsten disulfide. *Nature* **360**, 444–446 (1992).
17. Zak, A. et al. Scaling up of the WS₂ nanotubes synthesis. *Fuller. Nanotub. Carbon Nanostruct.* **19**, 18–26 (2010).
18. Zhang, C. et al. High-performance photodetectors for visible and near-infrared lights based on individual WS₂ nanotubes. *Appl. Phys. Lett.* **100**, 243101 (2012).
19. Qin, F. et al. Superconductivity in a chiral nanotube. *Nat. Commun.* **8**, 14465 (2017).
20. Yadgarov, L. et al. Strong light–matter interaction in tungsten disulfide nanotubes. *Phys. Chem. Chem. Phys.* **20**, 20812–20820 (2018).
21. Yuan, H. T. et al. Generation and electric control of spin-valley-coupled circular photogalvanic current in WSe₂. *Nat. Nanotechnol.* **9**, 851–857 (2014).
22. Freitag, M., Low, T., Xia, F. N. & Avouris, P. Photoconductivity of biased graphene. *Nat. Photon.* **7**, 53–59 (2013).
23. Quereda, J. et al. Symmetry regimes for circular photocurrents in monolayer MoSe₂. *Nat. Commun.* **9**, 3346 (2018).
24. Tauc, J. Generation of an emf in semiconductors with nonequilibrium current carrier concentrations. *Rev. Mod. Phys.* **29**, 308–324 (1957).
25. Yang, M. M., Kim, D. J. & Alexe, M. Flexo-photovoltaic effect. *Science* **360**, 904–907 (2018).
26. Yuan, Y. B., Xiao, Z. G., Yang, B. & Huang, J. S. Arising applications of ferroelectric materials in photovoltaic devices. *J. Mater. Chem. A Mater. Energy Sustain.* **2**, 6027–6041 (2014).
27. Morimoto, T. & Nagaosa, N. Topological nature of nonlinear optical effects in solids. *Sci. Adv.* **2**, e1501524 (2016).
28. Morimoto, T. & Nagaosa, N. Topological aspects of nonlinear excitonic processes in noncentrosymmetric crystals. *Phys. Rev. B* **94**, 035117 (2016).
29. Fregoso, B. M., Morimoto, T. & Moore, J. E. Quantitative relationship between polarization differences and the zone-averaged shift photocurrent. *Phys. Rev. B* **96**, 075421 (2017).
30. Rangel, T. et al. Large bulk photovoltaic effect and spontaneous polarization of single-layer monochalcogenides. *Phys. Rev. Lett.* **119**, 067402 (2017).
31. Nakhmanson, S. M., Calzolari, A., Meunier, V., Bernholc, J. & Nardelli, M. B. Spontaneous polarization and piezoelectricity in boron nitride nanotubes. *Phys. Rev. B* **67**, 235406 (2003).
32. Král, P., Mele, E. J. & Tomanek, D. Photogalvanic effects in heteropolar nanotubes. *Phys. Rev. Lett.* **85**, 1512–1515 (2000).
33. Pearce, A. J., Mariani, E. & Burkard, G. Tight-binding approach to strain and curvature in monolayer transition-metal dichalcogenides. *Phys. Rev. B* **94**, 155416 (2016).

Acknowledgements We thank M. Kuehne, H. Isobe, M. Nakamura, N. Ogawa and D. Zhao for discussions; S. Göres Y. Stuhlhofer, J. Geurs and Y. Kim for technical assistance; and A. Oiwa, B. Zhang, K. von Klitzing, Y. Tokura, M. Kawasaki and N. Nagaosa for suggestions. We acknowledge financial support from the Japan Society for the Promotion of Science (to Y.J.Z., M.O. and R.S. through the research fellowship program for young scientists; M.O. through the Advanced Leading Graduate Course for Photon Science; T.I. through the Challenging Research (Exploratory) (no. JP17K18748), the ‘Topological Materials Science’ (no. JP18H04216) KAKENHI on Innovative Areas, and Scientific Research (B) (no. JP19H01819); and Y.I. through a grant-in-aid for specially promoted research (no. 25000003) and Scientific Research (A) (no. JP19H00653)). We also acknowledge financial support from the Israel Science Foundation (ISF; to A.Z. and R.T.; no. 330/16 and 339/18); the Pazy Foundation of Israel (to A.Z.); the H. Perlman and the Irving and Azelle Waltcher Foundations in honour of M. Levy (to R.T.); the H. Perlman family foundation (to R.T.; no. 720821); and the Graphene Flagship (to J.H.S.).

Author contributions R.S. synthesized single crystals of WS₂. A.Z. synthesized and characterized WS₂ nanotubes, including carrying out X-ray diffraction and TEM measurements. Y.J.Z., M.O., F.Q. and R.S. fabricated devices and performed photovoltaic response measurements. Y.J.Z. performed SKPM measurements. Y.J.Z., T.I., A.Z., R.T., J.H.S. and Y.I. were the main writers of the manuscript. All authors contributed to improving the manuscript.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-019-1303-3>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to Y.J.Z.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

METHODS

Device fabrication. The chemically grown multiwall WS₂ nanotubes¹⁷ were dispersed on a SiO₂/Si⁺⁺ substrate. Isolated nanotubes were selected under an optical microscope. Single crystals of WS₂ were grown by chemical vapour transport³⁴. Bilayers and monolayers were mechanically exfoliated using an adhesive tape. Subsequently, the flakes were transferred to a SiO₂/Si⁺⁺ substrate. Their layer number was identified by the contrast in the optical micrographs. The electrodes were patterned using electron-beam lithography with poly(methylmethacrylate) (PMMA) as the resist. After developing the PMMA, chromium (5 nm) and gold (200 nm) were deposited as the electrodes via evaporation.

Photovoltaic response measurements. All electrical measurements were performed under high-vacuum conditions in an optical cryostat (MicrostatHe from Oxford Instruments) retrofitted with electrical feedthroughs. Samples were illuminated either with a helium–neon laser (wavelength 632.8 nm) or with one of two diode lasers (532 nm or 730 nm), with a laser spot size, as determined from a Gaussian fit to the laser profile, of approximately 2.5 μm (632.8 nm), 2.1 μm (532 nm) or 1.2 μm (730 nm). The laser beam was focused on samples at normal incidence through a ×50 Olympus objective. The tight focusing of the laser beam does not qualitatively affect the experimental results (see below). The incident laser power was monitored with a power-meter positioned behind the objective. The polarization characteristics were modulated by combining a Glan–Taylor prism, a λ/2 plate and a λ/4 plate. Data in the main text were acquired using a source measure unit (Keithley 2612B) under continuous illumination. Representative measurements were also repeated using lock-in techniques²² (see below). For nanotube devices, the input terminals were connected to the measurement device to ensure a positive signal of the I_{sc} induced by the BPVE.

Crystal and electronic structure of WS₂ nanotubes. The results of nearly all of our efforts to characterize the batch of nanotubes used here have already been reported. X-ray diffraction (Extended Data Fig. 1a) revealed that, locally, the structure of each layer resembles that of a monolayer with one W atom surrounded by six S atoms forming a triangular prism¹⁹. The Raman spectra of these nanotubes exhibit E_{2g} and A_{1g} vibrational modes similar to those of the bulk³⁵. The chirality of the nanotubes has been examined by high-resolution TEM³⁶ and electron diffraction³⁷ (Extended Data Fig. 1b). The layers that make up the multiwall nanotube have different chirality, and in general a single multiwall nanotube device is composed of zigzag, armchair and chiral single-wall tubes. Analogous to boron nitride nanotubes³⁸, their point groups are C_{2nh}, C_{2nv}, and C_N, respectively, and so both the zigzag and the chiral single-wall tubes are polar in nature and exhibit non-centrosymmetric symmetry. In all investigated devices, chiral single-wall tubes were present¹⁹. Hence, the multiwall WS₂ nanotube devices will overall also be non-centrosymmetric and polar, because a cancellation among the tubes with different diameter to restore a non-polar or centrosymmetric symmetry cannot be expected.

We determined the tube diameter of our WS₂ nanotube devices from a height measurement using atomic force microscopy (AFM; Extended Data Fig. 1c, d). It varies between tubes, but for all of them it is of the order of 10² nm. Because of their large diameter, the electronic structure of the nanotubes is anticipated to be similar to that of bulk material³⁹. Indeed, the direct gap (A- and B-exciton) energies of the investigated WS₂ nanotubes are comparable to those of bulk⁴⁰.

Dependence of the BPVE on the position of the laser spot. Extended Data Fig. 2a shows the dependence of I_{sc} on the position of the laser spot in WS₂ nanotube device 2. Similar to the first device (Fig. 2c), I_{sc} reaches a maximum when the centre of the nanotube is illuminated. In addition, a small peak is observed when the laser spot illuminates the boundary between the contact and the nanotube. This can be understood by considering multiple contributions from the BPVE, the Schottky barrier photovoltaic effect, and/or the photothermal effect^{21,22}. The BPVE should be largest when the laser spot is far away from the contacts. It should also remain constant—apart from variations caused by sample inhomogeneities—even if the laser spot is moved, provided that the contacts are not illuminated. As the laser approaches one of the contacts, part of the laser light will be reflected and as a result the total absorbed power by the nanotube is effectively reduced. This leads to a reduction in the BPVE. By contrast, a photovoltaic response within the Schottky barrier appears only when the laser spot is close to one of the contacts. The sign of this contribution is opposite for the two contacts. The magnitude is not necessarily the same as the quality and details of the sample/contact interface may vary substantially. This is evidenced in the asymmetric I – V characteristic of the device in the dark (Extended Data Fig. 2b). The relatively large I_{sc} at the left electrode (Extended Data Fig. 2a) reflects the fact that the photocurrents originating from the BPVE and the Schottky barrier photovoltaic effect have the same sign. At the right electrode, the signs of these two photocurrents are instead opposite. The absence of a clear minimum indicates that the Schottky barrier effect at this contact is rather small. For the sake of completeness, we note that the photothermal effect contributes in a similar fashion to the overall photocurrent as the Schottky barrier photovoltaic effect²¹.

Variation in the BPVE in WS₂ nanotubes. We observed the BPVE in five different WS₂ nanotube devices. The maximum I_{sc} , the excitation wavelength and the laser

power (P_{laser}) for each device are listed in Extended Data Table 1. These values were obtained by positioning the laser spot near the centre of the nanotube. Laser excitation with a wavelength of 632.8 nm tends to produce larger photocurrents than does a wavelength of 532 nm. This wavelength dependence follows the difference in the absorption coefficient²⁰. The enhanced response at 632.8 nm also agrees with the results summarized in Fig. 3d, where P_{laser} dependence of I_{sc} on a single device was recorded for three lasers with different wavelengths.

Light absorption in WS₂ nanotubes and monolayers. In order to estimate the portion of the light absorbed by the nanotube, we calculate the area of overlap between the incident laser beam and the nanotube. These calculations assume that the laser-beam profile has a Gaussian distribution and that the absorption coefficient is 100%. Taking into account the tube diameter (Extended Data Fig. 1d), we find an upper bound for the amount of absorbed light of about 11.4% and 5.7% of the total incident laser power for nanotube devices 1 and 2, respectively. If the absorption coefficient is less than 100%, the absorbed light intensity will be less accordingly. The width of the monolayer used in the measurements of Fig. 2b is 2 μm. About 89% of the incident light hits the monolayer area. The absorption coefficient at a wavelength of 632.8 nm, experimentally determined previously⁴⁰, is 7%. Hence, the WS₂ monolayer absorbs approximately 6.2% of the total incident laser power. **Temperature dependence of the BPVE.** Although the overall shape of the absorption spectrum remains similar with temperature, it does undergo a blue shift with decreasing temperature as a result of the temperature-dependent band gap^{11,41} ($E_g(T)$). The latter can be estimated from the following equation⁴²:

$$E_g(T) = E_g(0) - S <\hbar\omega> \left[\coth\left(\frac{\hbar\omega}{2k_B T}\right) - 1 \right]$$

Here, S is a dimensionless coupling constant, k_B is the Boltzmann constant, T is the temperature, and $<\hbar\omega>$ is an average phonon energy. By adopting parameters from the direct gap in a WS₂ bilayer ($S = 2.7$ and $<\hbar\omega> = 10.8$ meV)⁴³, we obtained $E_g(150\text{ K}) = E_g(300\text{ K}) + 70$ meV. If we assume that the whole absorption spectrum of a WS₂ nanotube at room temperature²⁰ shifts by +70 meV when reducing the device temperature to 150 K, the absorption coefficient at a wavelength of 632.8 nm will drop down to approximately 80% of its value at 300 K. In the WS₂ nanotube device, the photocurrent is reduced by more than one order of magnitude when cooling down from 300 K to 150 K (Fig. 3c). Therefore, the temperature-dependent absorption coefficient can be excluded as the primary source of the reduction in the photocurrent upon cooling the device.

An equivalent circuit model for BPVE devices includes contact resistances, which effectively reduce I_{sc} compared with the net photocurrent induced by the BPVE⁷. The contact resistances in general increase dramatically when the device is cooled down. This may account for the quick decrease in I_{sc} shown in Fig. 3c. In general, we note that a more refined model incorporating weaker but non-negligible contributions from, for example, the photogalvanic and photon drag effects (as discussed below), whose temperature dependencies are still unknown, is desirable.

Influence of oblique incidence. A finite zero-bias photocurrent in 2D TMDs arises at an oblique incidence of the laser light owing to the so called photogalvanic and photon drag effects^{21,23}. For a curved geometry like a nanotube, parts of the device will always be illuminated at oblique incidence (Extended Data Fig. 3a), and hence the observed signal may, at least to some extent, be similar to that reported for these 2D TMDs at oblique incidence. To evaluate this contribution, we recorded I_{sc} with an experimental configuration identical to that used previously for 2D TMDs^{21,23}. The incident laser beam is first linearly polarized in the x – z plane (perpendicular to the tube axis, defined as the y -axis). A λ/4 plate is then inserted to modulate the polarization. The parameter φ describes the angle between the fast axis of the λ/4 plate and the direction of the electric field of the linearly polarized laser beam. The photogalvanic and the photon drag effects can be separated by recording the φ dependence of the photocurrent^{21,23}. A typical example of the observed behaviour in WS₂ nanotubes is illustrated in Extended Data Fig. 3b. The data set can be fitted to the following expression, which contains φ -dependent terms that describe both the circular and the linear photogalvanic and photon drag effects, as well as a φ -dependent term relevant to the BPVE:

$$I_{sc}(\varphi) = \sin(2\varphi + \varphi_0) + L_G \sin(4\varphi + \varphi_0) + L'_D \cos(4\varphi + \varphi_0) + B'$$

Here, C is the amplitude of the circular photogalvanic and photon drag effects; L_G describes the magnitude of the linear photogalvanic effect; and B' is the average of the BPVE. The BPVE may also generate an oscillatory component following the same $\cos(4\varphi)$ dependence as the linear photon drag effect owing to the anisotropy of the absorption coefficient (Fig. 2d). L'_D thus describes the sum of this oscillatory part of the BPVE and the linear photon drag effect. In order to compensate for some misalignment in the optics, φ_0 has been added as an additional fit parameter. The best fit to the data (red solid line in Extended Data Fig. 3b) is obtained for $C = -0.99$ nA, $L_G = 0.77$ nA, $L'_D = -2.1$ nA, $B' = 8.7$ nA and $\varphi_0 = 10^\circ$.

The contribution from the BPVE to I_{sc} is much larger than the contribution from the photogalvanic and photon drag effects.

Relevance of the Dember effect. The Dember effect is caused by a difference in the mobility of electrons and holes⁴¹. As the electron mobility and hole mobility of WS₂ are indeed different⁴⁴, the Dember effect may emerge and generate two electromotive forces pointing towards the two electrodes in our WS₂ nanotube devices. Ideally, these two electromotive forces would compensate each other. This cancellation can be suppressed when an inhomogeneous photocurrent generation occurs owing to an asymmetric distribution of the incident light intensity and/or an intrinsic inhomogeneity in the sample. In our set-up, the laser-spot profile is symmetric and fits a Gaussian function well. Therefore, we can exclude the first scenario of an asymmetric light-intensity profile.

The homogeneity of the WS₂ nanotubes can be assessed by measuring the surface potential variation along the nanotube. Extended Data Fig. 4a, b show representative AFM and scanning Kelvin probe microscopy (SKPM) images of a WS₂ nanotube device in the dark, respectively. The surface potential profiles along three different WS₂ nanotubes are displayed in Extended Data Fig. 4c. Extended Data Fig. 4d illustrates how the photovoltaic effect owing to the Dember effect would vary with the position of the laser spot when the sample is inhomogeneous and composed of two (left), four (centre) or many (right) domains with two different electrical characteristics^{24,45,46}. The key observation is an intimate correlation between the photovoltaic signal and the spatial distribution of the domains. Extended Data Fig. 4c suggests that the scenario illustrated in the right panel of Extended Data Fig. 4d should apply to our WS₂ nanotube devices. In this case, the remnant Dember effect will be small and the photovoltaic signal should fluctuate around zero. This is qualitatively inconsistent with the experimentally recorded data (Fig. 2c).

We also address the Dember effect quantitatively. According to ref.²⁴, the maximum photovoltage induced by the Dember effect is given by:

$$V_{\max} = \frac{k_B T}{e} \left(\frac{\mu_e - \mu_h}{\mu_e + \mu_h} \right) \ln \frac{1 + \sigma_{ph}/\sigma_{0,x2}}{1 + \sigma_{ph}/\sigma_{0,x1}} \approx \frac{k_B T}{e} \left(\frac{\mu_e - \mu_h}{\mu_e + \mu_h} \right) \ln \frac{\sigma_{0,x1}}{\sigma_{0,x2}}$$

Here, σ_{ph} refers to the photoconductivity and $\sigma_{0,x}$ to the dark conductivity at the left and right boundaries ($x1$ and $x2$) of the illuminated region. The electron and hole mobility are denoted by μ_e and μ_h , respectively. The approximation is valid for the case of strong illumination (where σ_{ph} is much greater than σ_0)²⁴. The I - V characteristics in Fig. 3b imply that this condition is satisfied in our measurement at the maximum P_{laser} , where V_{oc} reaches a value of approximately 0.4 V. Under the assumption that this photovoltage is generated by the Dember effect, the sample must have at least one domain boundary where the dark conductivities on either side of the boundary (positions $x1$ and $x2$) satisfy the inequality:

$$0.4 \leq |V_{\max}| \leq \frac{k_B T}{e} \left| \ln \frac{\sigma_{0,x1}}{\sigma_{0,x2}} \right|$$

At room temperature (300 K), this leads to:

$$\frac{\sigma_{0,x1}}{\sigma_{0,x2}} \geq 5.24 \times 10^6 \text{ or } \frac{\sigma_{0,x2}}{\sigma_{0,x1}} \geq 5.24 \times 10^6$$

Hence, the dark conductivity at positions $x1$ and $x2$ must differ by more than six orders of magnitude. Extended Data Fig. 4c indicates that the variation in the potential along the nanotubes is at best 100 meV. According to previous studies of WS₂ nanotube field-effect transistors with electric double layers as gate dielectrics^{19,35}, the dark conductivity in WS₂ nanotubes caused by a change in the Fermi energy of 100 meV is at best two orders of magnitude. This is four orders of magnitude less than the difference required to account for the observed V_{oc} based on the Dember effect.

Anisotropic absorption in WS₂ nanotubes. As shown in Fig. 2d, the absorption coefficient of the WS₂ nanotubes may depend on direction of the linear polarization of the light. Strictly speaking, this anisotropy also needs to be considered when discussing the photovoltaic response to circularly polarized light, because the incident beam will acquire a component with opposite helicity. However, because most of the light remains of the same circular polarization direction, the photovoltaic response under light illumination with circular polarization σ_+ (σ_-) still originates predominantly from the carriers excited at the K (K') point. Therefore, the conclusions drawn remain valid despite the anisotropy in the absorption coefficient.

Influence of device design on BPVE external output. The external output produced by the BPVE does not depend only on the physical properties of the material used⁴⁻¹⁰. The device configuration^{8,47-50} and its parameters also affect the output. For instance, I_{sc} induced by the BPVE is typically orders of magnitude larger in a vertical device than in a lateral device, because of the improved charge-collection efficiency^{8,47,48}. The Schottky barrier also contributes to the output and can improve the overall efficiency⁴⁹. Recently, the photocurrent in BiFeO₃ was

dramatically improved by introducing an AFM tip to locally enhance the charge-collection efficiency⁵⁰. Extended Data Fig. 5 summarizes the efficiency improvement achieved in each material by altering the device configuration. The photocurrent generated by the BPVE in our WS₂ nanotube devices is comparable with, or even larger than, that obtained for devices fabricated from other materials, after some optimization of the device configuration. The difference would be even larger if we keep in mind that the incident laser beam only has a roughly 10% overlap with the WS₂ nanotube devices. As no effort has been made to improve the external output of our WS₂ nanotube devices, there is room for further improvement.

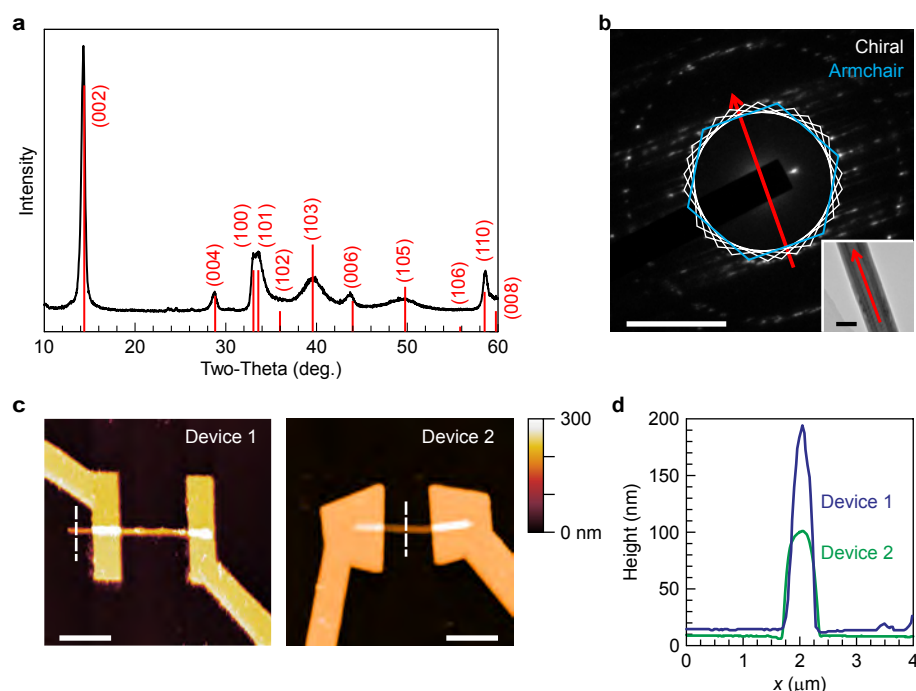
Measurements under different focusing conditions. We also recorded the photovoltaic effect in WS₂ nanotube device 2 with a $\times 10$ objective. For the laser with a wavelength of 532 nm, the spot size is 45.2 μm —about one order of magnitude larger than the channel length. Hence under these conditions the device is, in essence, uniformly illuminated. The results are summarized in Extended Data Fig. 6. A photovoltaic effect is clearly observed (Extended Data Fig. 6a), and I_{sc} also shows $\cos(2\theta)$ dependence upon the direction of linear polarization (Extended Data Fig. 6b). The power dependence (Extended Data Fig. 6c) reveals a kink near 10 W cm⁻², consistent with the crossover from a linear to a square-root dependence observed with focused laser-spot illumination (Fig. 3d). It is reasonable that I_{sc} increases faster than a square-root dependence, because I_{sc} inevitably includes a contribution from the Schottky barrier photovoltaic effect, which increases linearly within the investigated power region¹⁸. As highlighted in Extended Data Fig. 6d, V_{oc} also varies with temperature for illumination with a widened laser beam.

AC measurements. We repeated some of the photocurrent measurements with the help of lock-in techniques (Extended Data Fig. 7). The experimental set-up is identical to that in ref.²². Extended Data Fig. 7a displays the magnitude and phase of the AC photocurrent as a function of the position of the laser spot, whereas Extended Data Fig. 7b plots the dependence of the photocurrent amplitude and phase as a function of the linear polarization direction. The amplitude reaches its maximum when the laser spot is located away from the contacts (Extended Data Fig. 7a) and when the incident light is linearly polarized along the tube axis (Extended Data Fig. 7b). These dependencies are consistent with the DC measurements shown in Extended Data Fig. 2a and the blue symbols of Fig. 2d. The phase remains almost constant in both measurements. The voltage dependence of the magnitude and phase of the AC photocurrent is plotted in Extended Data Fig. 7c. The dramatic phase change of 180° when the photocurrent approaches 0 nA reflects the sign reversal of the photocurrent in DC measurements.

Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

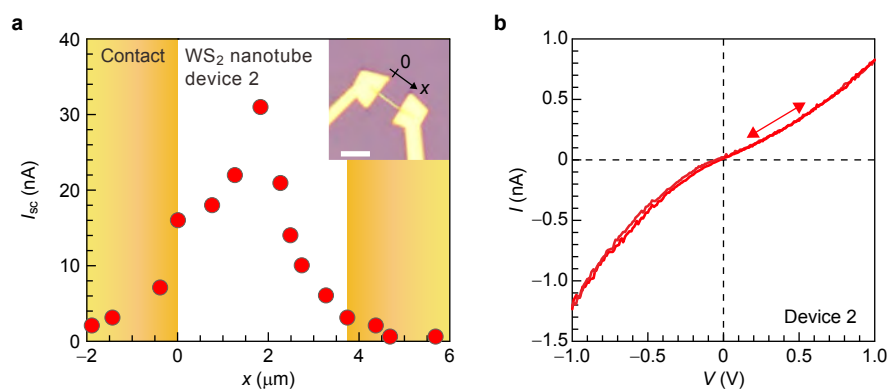
34. Shi, W. et al. Superconductivity series in transition metal dichalcogenides by ionic gating. *Sci. Rep.* **5**, 11534 (2015).
35. Zhang, Y. J. et al. Optoelectronic response of a WS₂ tubular p - n junction. *2D Mater.* **5**, 035002 (2018).
36. Bar Sadan, M., Houben, L., Enyashin, A. N., Seifert, G. & Tenne, R. Atom by atom: HRTEM insights into inorganic nanotubes and fullerene-like structures. *Proc. Natl Acad. Sci. USA* **105**, 15643–15648 (2008).
37. Chen, Y. H., Deniz, H. K. & Qin, L. C. Accurate measurement of the chirality of WS₂ nanotubes. *Nanoscale* **9**, 7124–7134 (2017).
38. Alon, O. E. Symmetry properties of single-walled boron nitride nanotubes. *Phys. Rev. B* **64**, 153408 (2001).
39. Ghorbani-Asl, M. et al. Electromechanics in MoS₂ and WS₂: nanotubes vs. monolayers. *Sci. Rep.* **3**, 2961 (2013).
40. Bernardi, M., Palummo, M. & Grossman, J. C. Extraordinary sunlight absorption and one-nanometer-thick photovoltaics using two-dimensional monolayer materials. *Nano Lett.* **13**, 3664–3670 (2013).
41. Seeger, K. *Semiconductor Physics: An Introduction* 2nd edn (Springer-Verlag, 1982).
42. O'Donnell, K. P. & Chen, X. Temperature dependence of semiconductor band-gaps. *Appl. Phys. Lett.* **58**, 2924–2926 (1991).
43. He, Z. Y. et al. Layer-dependent modulation of tungsten disulfide photoluminescence by lateral electric fields. *ACS Nano* **9**, 2740–2748 (2015).
44. Braga, D., Lezama, I. G., Berger, H. & Morpurgo, A. F. Quantitative determination of the band gap of WS₂ with ambipolar ionic liquid-gated transistors. *Nano Lett.* **12**, 5218–5223 (2012).
45. Trousil, Z. Bulk photo-voltaic phenomenon. *Czech. J. Phys.* **6**, 96–98 (1956).
46. Frank, H. Lichte elektrische Messung des inneren elektrischen Feldes in inhomogenen Halbleitern. *Czech. J. Phys.* **6**, 433–441 (1956).
47. Ichiki, M. et al. Photovoltaic effect of lead lanthanum zirconate titanate in a layered film structure design. *Appl. Phys. Lett.* **84**, 395–397 (2004).
48. Cao, D. W. et al. High-efficiency ferroelectric-film solar cells with an n -type Cu₂O cathode buffer layer. *Nano Lett.* **12**, 2803–2809 (2012).
49. Zenkevich, A. et al. Giant bulk photovoltaic effect in thin ferroelectric BaTiO₃ films. *Phys. Rev. B* **90**, 161409 (2014).
50. Alexe, M. & Hesse, D. Tip-enhanced photovoltaic effects in bismuth ferrite. *Nat. Commun.* **2**, 256 (2011).
51. Zak, A., Sallacan-Ecker, L., Margolin, A., Genut, M. & Tenne, R. Insight into the growth mechanism of WS₂ nanotubes in the scaled-up fluidized-bed reactor. *Nano* **4**, 91–98 (2009).



Extended Data Fig. 1 | Characterization of WS₂ nanotube devices.

a, The black trace shows the results of an X-ray diffraction analysis of a WS₂ nanotube. The red bars indicate the simulated peak positions for a 2H-type bulk WS₂ material. The small shift in the (002), (004) and (006) peaks for the WS₂ nanotube is also consistent with previous work⁵¹ indicating that the distance between layers in the nanotubes is slightly larger than that in bulk WS₂. **b**, Electron-diffraction pattern for a WS₂ nanotube. White and blue hexagons have been included to indicate the

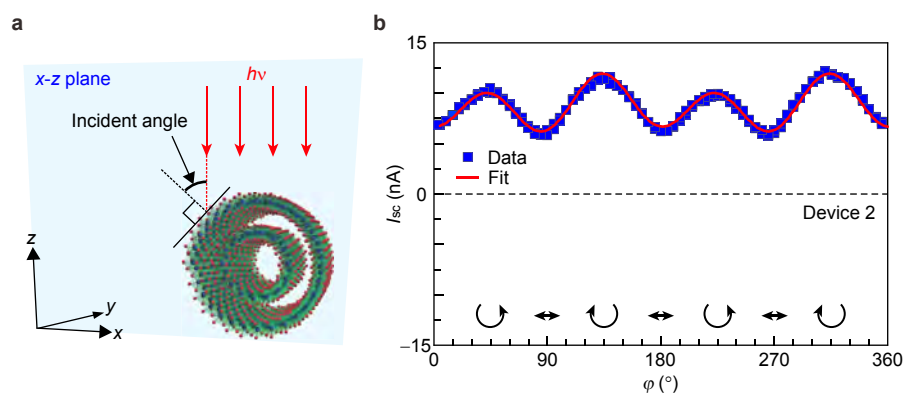
differing chirality of each layer making up the nanotube wall¹⁹. The white scale bar represents 5 nm⁻¹. The red arrow marks the orientation of the nanotube, as determined from the TEM image in the inset (black scale bar, 50 nm). **c**, Colour rendition of the height map recorded by AFM on devices 1 and 2. Scale bars represent 4 μm. **d**, Height profiles of both tubes along the dashed lines in panel **c**. The horizontal axis (*x*) represents the position along these lines. The diameter of the first nanotube (device 1) is 180 nm; for device 2, the diameter is 90 nm. **a** and **b** are adapted from ref. ¹⁹.



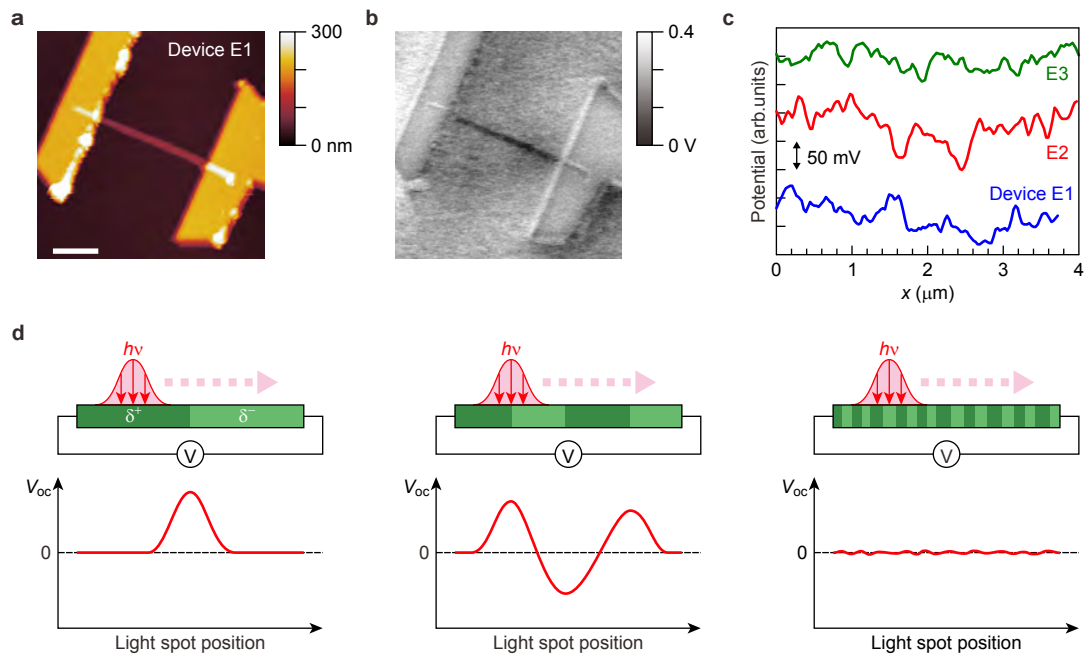
Extended Data Fig. 2 | Photovoltaic response of WS_2 nanotube device 2.

a, Dependence of I_{sc} on the position of the laser spot. The scale bar in the optical micrograph of the device represents $4\ \mu\text{m}$. The horizontal

axis (x) shows the distance between the laser spot and one of the electrodes (see optical micrograph, inset). **b**, DC I - V characteristics in the dark at voltages up to $\pm 1\ \text{V}$.

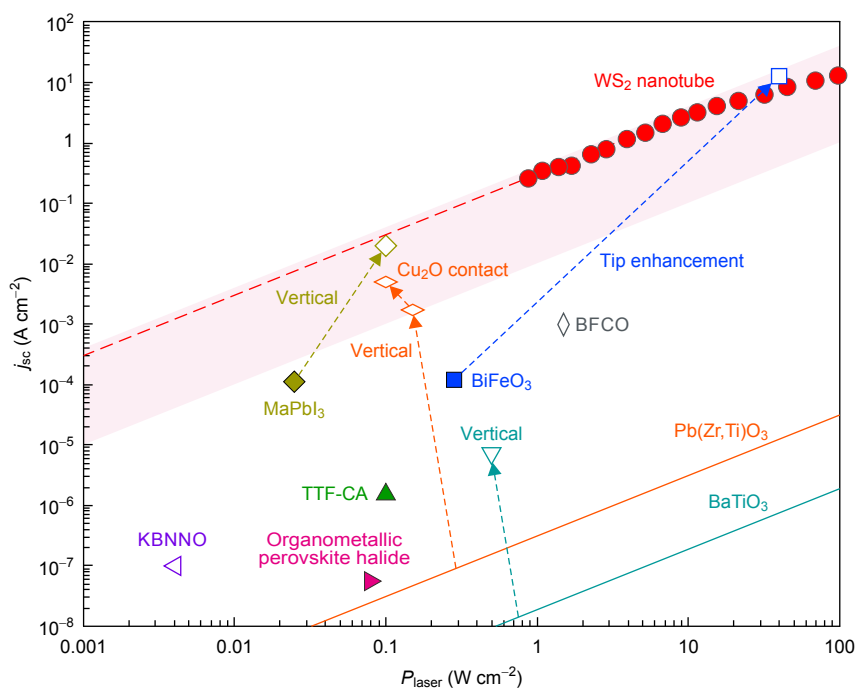


Extended Data Fig. 3 | Effect of oblique incidence of the laser light on the photocurrent response of a WS₂ nanotube device. a, Illustration of the oblique incidence condition in WS₂ nanotubes. **b,** Dependence of the photocurrent on the polarization characteristics of the incident laser beam.



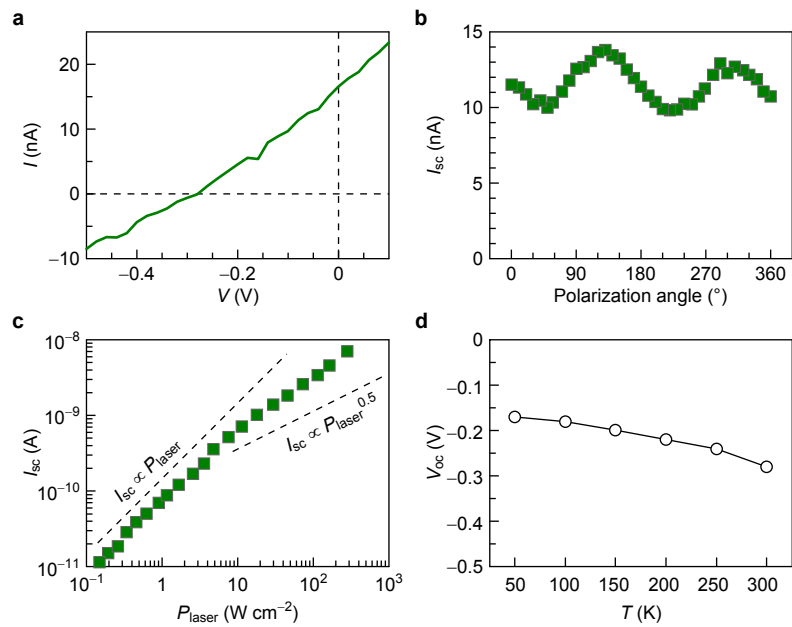
Extended Data Fig. 4 | AFM and SKPM of WS₂ nanotubes, and influence of domain number and size on the net Dember photovoltaic effect. a, Colour rendition of a height map recorded by AFM in the dark. The scale bar represents 2 μm . **b,** Greyscale map of the surface potential recorded with a scanning Kelvin probe microscope in the dark. The scanned area is identical to that in **a**. The sample contacts are

both connected to ground during the measurement. **c,** Line scans of the surface potential along three different WS₂ nanotubes. **d,** Expected spatial dependence of the photovoltaic signal caused by the Dember effect for three different domain configurations. Domains with different electrical characteristics are coloured differently and marked with δ^+ (dark green) or δ^- (light green).



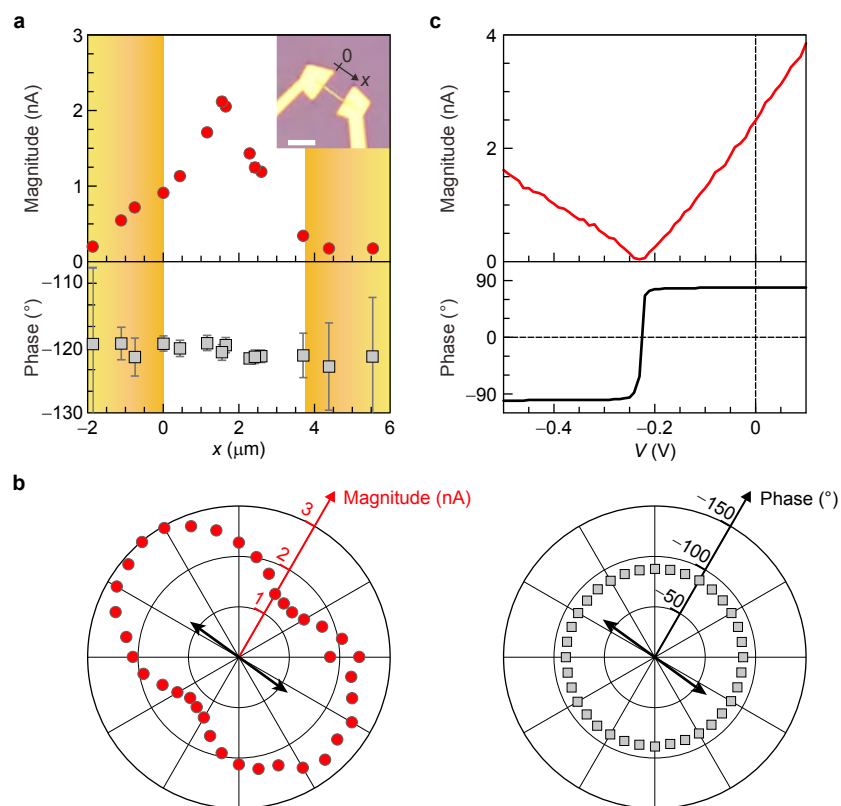
Extended Data Fig. 5 | The BPVE in various materials. The short-circuit current density, j_{sc} , is plotted against laser power for bulk ferroelectric materials and WS₂ nanotubes. Filled symbols are used for data points obtained in a lateral device configuration, whereas open symbols refer to data points obtained on other, vertical device geometries. References are

as follows: KBNNO⁴; TTF-CA⁵; Pb(ZrTi)O₃, lateral⁶, vertical⁴⁷, contact with copper oxide⁴⁸; BaTiO₃, lateral⁶, vertical⁴⁹; BiFeO₃, lateral⁷, tip enhancement⁵⁰; MaPbI₃, lateral and vertical⁸; organometallic perovskite halide⁹; BFCO¹⁰.



Extended Data Fig. 6 | Photovoltaic response of WS₂ nanotube device 2 under uniform illumination. The laser wavelength was 532 nm. **a**, I - V characteristics recorded with a $\times 10$ objective. **b**, Dependence of I_{sc} on the linear polarization angle of the incident laser beam. **c**, Laser power

dependence of I_{sc} . The dashed lines correspond to a linear and a square-root power dependence and serve as guides to the eye. **d**, Temperature dependence of V_{oc} . The laser power density was set to $8.29 \times 10^2\ W\ cm^{-2}$ for **a**, **b** and **d**. The data in **a**-**c** were recorded at 300 K.



Extended Data Fig. 7 | Lock-in measurements of the BPVE in WS₂ nanotube device 2. The incident laser beam was periodically modulated at a frequency of 133 Hz with the help of a chopper. The excitation wavelength and laser power were 632.8 nm and $6.32 \times 10^3 \text{ W cm}^{-2}$, respectively. **a**, Dependence of I_{sc} on the position of the laser spot. The scale bar in the optical micrograph of the device represents 4 μm . Both

magnitude and phase are plotted. The horizontal axis (x) is the distance between the laser spot and one of the electrodes (see optical micrograph). **b**, Dependence of the photocurrent on the linear polarization angle of the incident laser beam. The thick black arrows at the centre of the polar diagrams mark the orientation of the nanotube. **c**, Magnitude and phase of the AC photocurrent as a function of the applied voltage.

Extended Data Table 1 | Maximum short-circuit current in five WS₂ nanotube devices at the indicated wavelengths and laser intensities

	Device 1	Device 2	Device 3	Device 4	Device 5
Wavelength (nm)	632.8	632.8	632.8	532	532
Laser power (W cm ⁻²)	1.39×10^4	1.39×10^4	1.39×10^4	3.60×10^4	3.60×10^4
Maximum I_{sc} (nA)	15	31	16	0.5	10

Chiral twisted van der Waals nanowires

Peter Sutter^{1*}, Shawn Wimer¹ & Eli Sutter^{2*}

Van der Waals heterostructures with small misalignment between adjacent layers ('interlayer twist') are of interest because of electronic structure and correlation phenomena (such as superconductivity) that are determined by both the atomic lattice and long-range superlattice potentials arising in interlayer moiré patterns^{1–7}. Previously, such twisted heterostructures have involved a single planar interface between layers isolated by exfoliation and micromechanically stacked in the desired relative orientation^{1,8–12}. Here we demonstrate a class of materials—van der Waals nanowires of layered crystals—in which a tunable interlayer twist evolves naturally during synthesis. In vapour–liquid–solid growth, nanowires of germanium(ii) sulfide, an anisotropic layered semiconductor, crystallize with layering along the wire axis¹³ and have a strong propensity for forming axial screw dislocations. Nanometre-resolved electron diffraction shows that Eshelby twist, induced by a torque on the ends of a cylindrical solid due to the stress field of an axial dislocation^{14,15}, causes a chiral structure in the van der Waals nanowires. The in-plane germanium sulfide crystal axes progressively rotate along the wire, and germanium sulfide layers in adjacent turns of the helix naturally form a moiré pattern because of their interlayer twist. The axial rotation and the twist are tunable by varying the nanowire thickness. Combined electron diffraction and cathodoluminescence spectroscopy show the correlation between the interlayer twist and locally excited light emission that is due to progressive changes in the lattice orientation and in the interlayer moiré registry along the nanowires. The findings demonstrate a step towards scalable fabrication of van der Waals structures with defined twist angles, in which interlayer moiré patterns are realized along a helical path on a nanowire instead of a planar interface.

Three-dimensional (3D) crystalline semiconductor nanowires (such as Si, Ge or GaAs) have long attracted interest as a class of nanomaterials that can be synthesized with exceptionally high crystal quality through vapour–liquid–solid (VLS) growth, in which a nanoscale liquid 'catalyst' transports source material from the vapour phase to the growth front of a solid crystalline wire^{16,17}. Previous work on integrating layered crystals in nanowires includes VLS growth of semiconducting SnS (ref. ¹⁸), SnSe (ref. ¹⁹), GaSe (ref. ²⁰), In₂Se₃ (ref. ²¹) and Sb₂Se₃ (ref. ²²) wires, as well as Bi₂Te₃ and Bi₂Se₃ topological insulators²³. Around half of these reports showed van der Waals stacking perpendicular to the nanowire axis and half showed stacking parallel to the axis. Here we focus on layered GeS nanowires synthesized by Au-catalysed low-temperature VLS growth. Such nanowires are particularly interesting because of their anisotropic structure. A robust synthesis (see Methods), discussed in detail elsewhere¹³, provides GeS nanowires with different diameters and uniform structure. The wires crystallize with van der Waals layering (*c* axis) along their symmetry axis: that is, the *a* and *b* unit vectors spanning the individual anisotropic, covalently bonded layers lie in planes perpendicular to the wire (Fig. 1a, b). These structural characteristics are borne out in high-resolution images obtained by transmission electron microscopy (TEM) and high-angle annular dark-field scanning TEM (HAADF-STEM), which show lattice fringes with the GeS *c*-axis spacing of 1.06 nm across the nanowires, and by nanobeam electron diffraction (Fig. 1c). Nanometre-scale diffraction, which can probe the local crystal

orientation along the wires, shows another key property of the nanowire structure. Even if the crystal is locally aligned such that the electron beam projects along a low-index zone axis (here, approximately [1 0 0], Fig. 1c), the diffraction patterns rotate systematically away from this orientation as the beam is displaced along the nanowire. While the GeS *c* axis remains oriented along the wire, the (*a*, *b*) plane progressively rotates with a fixed helicity: that is, the nanowires are chiral. For the wire shown in Fig. 1, for instance, the cumulative effect is to rotate the (*a*, *b*) plane by about 21° over a distance of 1.8 μm.

Imaging and further diffraction analysis of the GeS nanowire crystallography clarify the origin of the observed unidirectional lattice rotation. In (S)TEM images, the registry of lattice fringes across the wires shows ubiquitous axial screw dislocations in the GeS van der Waals nanowires (Extended Data Fig. 1), consistent with the facile formation of screw dislocations or growth spirals in layered crystals²⁴, including GeS (ref. ²⁵) and SnS (ref. ²⁶). Examination of the spatially resolved diffraction data for nanowires with different diameters identifies two key characteristics (Fig. 2a): the cumulative lattice rotation increases linearly with displacement along the wires; and the rate of this increase shows a clear dependence on the nanowire diameter, with thinner wires rotating by larger angles per unit length. The mean rotation rate depends exponentially on the nanowire diameter across the size range accessed here (Fig. 2b), reaching about 0.1° per nanometre for nanowires below 40 nm in thickness. The presence of axial screw dislocations and the diameter-dependent twist are telltale signatures of Eshelby twist, induced by a torque on the ends of a cylindrical solid due to the stress field of an axial dislocation, as the origin of the chiral structure of the van der Waals nanowires. The Eshelby formalism¹⁴ relates the lattice rotation per unit length, $d\theta/dz$, to the cross-sectional area πR^2 of the cylinder and the Burgers vector *b* of the screw component of the axial dislocation: $d\theta/dz = b/(\pi R^2)$. Replotting our data as a function of the nanowire footprint confirms the realization of Eshelby twist in the dislocated van der Waals nanowires, similar to previous results for conventional 3D crystalline nanowires¹⁵, and it enables the Burgers vector of the axial screw dislocation to be extracted (Fig. 2c). From the diameter-dependent mean twist, we find a mean magnitude of the Burgers vector $|b| = (1.5 \pm 0.15)$ nm, corresponding to about 1.5 times the *c*-axis dimension of the GeS unit cell (computed, 1.077 nm (ref. ²⁷); experimental, 1.042 nm (ref. ²⁸)), which contains two weakly coupled GeS layers spaced by $L = 0.521$ nm. An alternative analysis using the full position-dependent diffraction dataset of Fig. 2a confirms this mean value of $|b|$ but provides more detailed insight into the distribution of Burgers vectors (Fig. 2d). It shows a preference for Burgers vectors of one and two *c*-axis unit cells (two or four GeS layer spacings, *L*, respectively; Fig. 2d, inset) and a suppression of small half-integer values. The overall Gaussian line shape of the histogram is consistent with the expected progressive reduction of the frequency for larger Burgers vectors.

The crystallographic analysis shows that the layered GeS nanowires are helical growth spirals with diameter-dependent crystal twist that predominantly involves one or two GeS unit cells—that is, two or four layers. The helicity implies the spontaneous formation of interlayer twist between the lattices in adjacent turns of the spiral, analogous to the twist achieved by artificial stacking of extended 2D or layered

¹Department of Electrical and Computer Engineering, University of Nebraska-Lincoln, Lincoln, NE, USA. ²Department of Mechanical and Materials Engineering, University of Nebraska-Lincoln, Lincoln, NE, USA. *e-mail: esutter@unl.edu; psutter@unl.edu

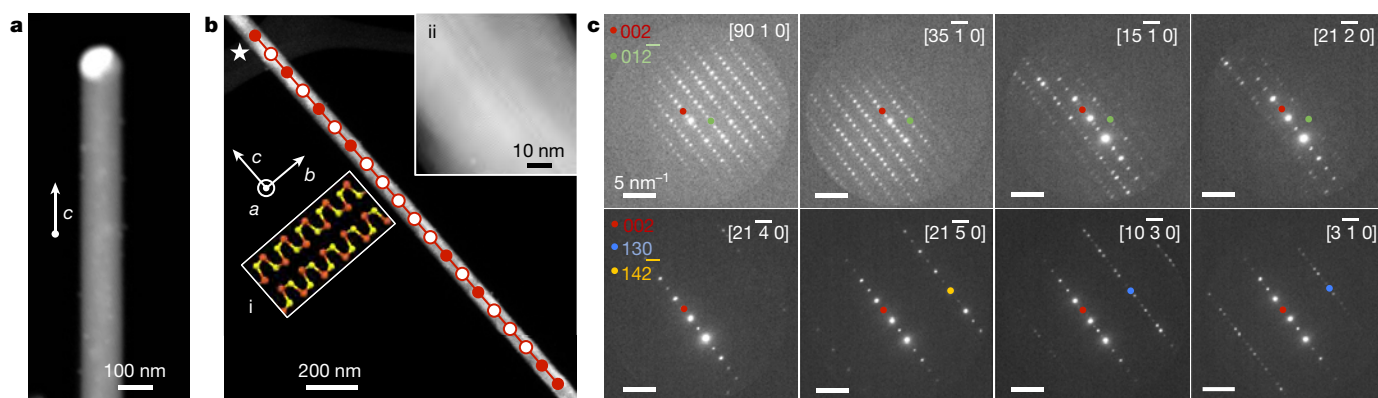


Fig. 1 | Twisted van der Waals nanowires. **a**, HAADF-STEM image of a layered GeS nanowire formed by Au-catalysed vapour–liquid–solid growth. **b**, Section of a GeS nanowire, analysed by nanobeam diffraction. Insets: **i**, layered GeS structure near the top left segment of the nanowire, marked by a star; **ii**, high-resolution STEM, showing layering with

1.06-nm periodicity along the wire axis. **c**, Selected nanobeam electron diffraction patterns at positions along the nanowire marked by red dots in **b**. Indices $[h\ k\ l]$ denote the actual zone axis of the nanowire, starting near $[1\ 0\ 0]$ and showing progressive twist of the lattice along the nanowire axis.

crystals^{1,7}. Similar to planar systems such as twisted bilayer graphene^{5,6}, this interlayer twist between identical 2D crystals generates a rotational moiré pattern that sets up a spatially varying superlattice potential at the twisted van der Waals interface (Fig. 2e). Owing to the large moiré period (Extended Data Fig. 2) and the way in which the twist moiré is projected onto a helical van der Waals interface along the nanowires (see Methods), direct probing of the moiré pattern by diffraction is not feasible. To detect signatures of the crystal rotation and interlayer moiré by means of the expected changes in electronic structure, we carried out cathodoluminescence (CL) spectroscopy excited locally by the focused electron beam in STEM. Results of such STEM-CL measurements are shown in Fig. 3.

Combined nanobeam electron diffraction and STEM-CL measured at room temperature show the correlation between the progressive

interlayer twist, here about 17° over 420 nm in length (Fig. 3a), and locally excited light emission from the nanowires. Series of CL spectra obtained along individual GeS nanowires (Fig. 3b) show that the cumulative twist along the wire is accompanied by synchronous changes in the intensity, width and centre wavelength of the GeS band-edge luminescence (Fig. 3c). Specifically, the centre wavelength shifts from about 580 nm to a minimum of 530 nm while the peak width increases from about 170 nm to 260 nm. Over the same distance, the amplitude of the Gaussian peak decreases by approximately 25%. Similar systematic changes in the spectral characteristics were obtained in STEM-CL at low temperature ($T = 110$ K; Extended Data Fig. 3). The centre wavelength of the nanowire luminescence is strongly blueshifted from the GeS bulk bandgap (1.60 eV)²⁹ and scales inversely with the wire diameter, suggesting a confinement effect (Extended Data Figs. 4, 5).

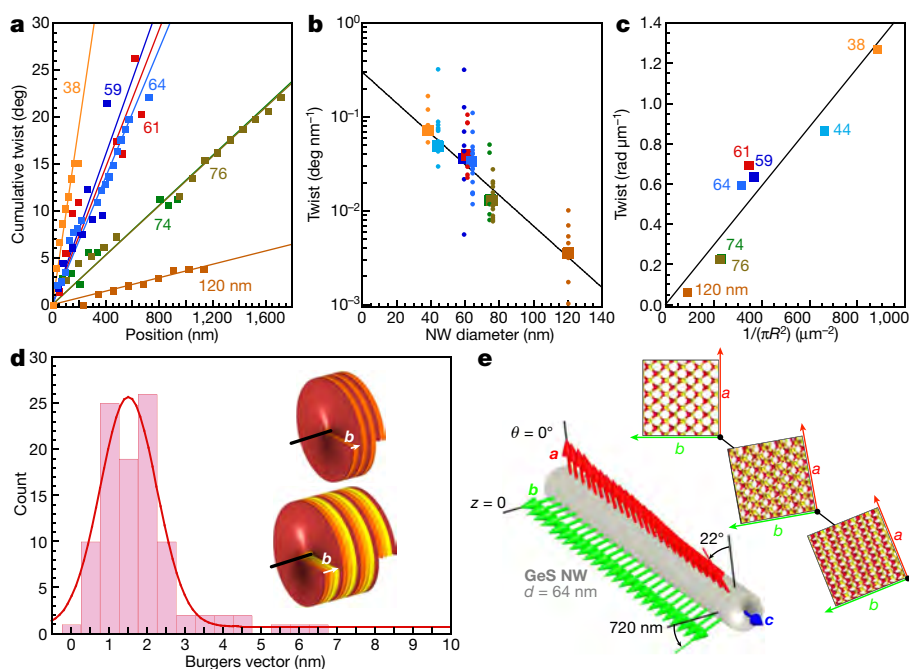


Fig. 2 | Eshelby twist of layered GeS nanowires. **a**, Measured cumulative twist angle along selected GeS nanowires and its dependence on wire diameter between 38 nm and 120 nm. **b**, Scatter plot of the twist per unit length for the nanowires shown in **a**. Large symbols, mean twist per nanometre with exponential dependence on nanowire (NW) diameter. **c**, Mean Eshelby twist of the wires shown in **a**, and linear fit giving a mean

Burgers vector $b = 1.5$ nm, that is, $\sim 3L$ (GeS layer spacings), of the axial screw dislocation. **d**, Histogram of Burgers vectors for the full dataset shown in **a**, showing preference for values of $2L$ and $4L$. Red line, Gaussian fit to the data. Inset, illustration of twisted van der Waals nanowires with $2L$ and $4L$ Burgers vectors. **e**, Schematic showing the rotation of the (a, b) crystal axes and the change in moiré registry along chiral GeS nanowires.

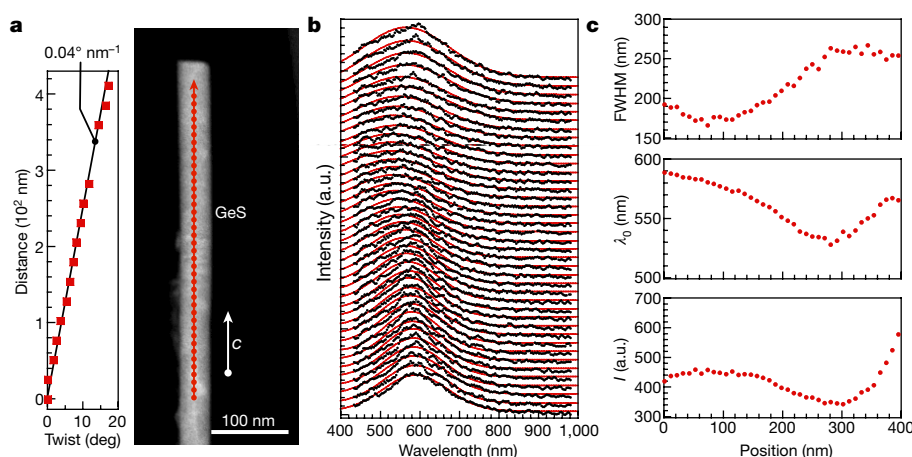


Fig. 3 | Optoelectronics of twisted GeS nanowires. **a**, HAADF-STEM image of a section of a GeS van der Waals nanowire and plot of cumulative twist determined by nanobeam diffraction analysis, showing progressive twist $d\theta/dz = 0.04^\circ \text{ nm}^{-1}$. **b**, Series of room-temperature cathodoluminescence spectra obtained at points marked in **a** (STEM

image, red dots). Red lines, Gaussian fits to the spectra. a.u., arbitrary units. **c**, Parameters of the Gaussian fits along the GeS nanowire: full-width at half-maximum (FWHM), centre wavelength (λ_0) and peak amplitude (I).

However, because the Eshelby twist also increases with decreasing diameter (Fig. 2), an effect of the twist on the bandgap cannot be excluded. The large peak width is the result of inhomogeneous line broadening. In STEM-CL, an entire chain of elementary processes, excitation, transport by diffusion or drift, and ultimately recombination is involved before photon emission³⁰. Control experiments show the redistribution of excited charge carriers due to electric-field-induced drift or asymmetric diffusion, probably because of the filling of trap states at the core of the axial screw dislocation (see Methods). Hence, despite the local electron-beam excitation, a finite section of the nanowire—covering a range of electronic structure, such as different bandgaps—is involved in light emission, which produces the observed broad luminescence peaks.

For chiral GeS nanowires, the position-dependent changes in the luminescence spectra could originate from two distinct effects: the progressive rotation of the van der Waals layers (that is, rotation of the (*a*, *b*) crystal axes) with respect to the exciting electron beam, or gradual changes in electronic structure due to a varying interlayer moiré registry. The electron-beam excitation and light collection in STEM-CL are largely isotropic and should not be sensitive to rotation of the anisotropic GeS lattice²⁹, but band-structure effects on the light emission for different incidence directions of the exciting electron beam cannot be ruled out entirely. Raman scattering and photoluminescence spectroscopy of GeS plates excited by linearly polarized light incident along the *c* axis show a strong polarization dependence of the Raman intensity of the major (B_{3g} , A_g) phonon modes, but only a small modulation of the photoluminescence intensity and minor shifts of the emission wavelength (less than 5 nm; Extended Data Fig. 6). Hence, the systematic changes in luminescence shown in Fig. 3 probably arise from modifications to the electronic structure due to a progressively varying interlayer moiré registry along the nanowires. An analysis of STEM-CL spectra obtained along extended (several micrometre) sections of GeS van der Waals nanowires shows spectral features correlated with the local moiré registry and thus supports the notion that changes in the twist moiré registry along the helical wires produce systematic variations in the electronic structure and optoelectronic properties (see Methods, Extended Data Figs. 7, 8).

Layered nanowires can therefore be expected to harbour emergent electronic phenomena found so far only in planar van der Waals heterostructures. Semiconductor wires, as shown here, promise modulated optoelectronic properties due to the twist superlattice⁴ and chiral light–matter interactions³¹ governed by both the helical structure and twist moiré. Realizing chiral wires from graphitic carbon would create a new platform for studying electron correlation effects of twisted

bilayer graphene^{5,6}. The 1D geometry has several attributes distinct from conventional 2D van der Waals stacks: chiral nanowires spontaneously generate an interlayer moiré through the Eshelby twist associated with axial screw dislocations; the twist angle and moiré periodicity are tunable by varying the nanowire diameter, which in turn can be selected by adjusting the size of the VLS catalyst³²; and the moiré registry varies systematically along a helical path instead of an extended planar interface. Because layered crystals readily form screw dislocations, chiral nanowires may be produced from different materials covering a wide range of electronic structure, provided that they crystallize with van der Waals stacking parallel to the wire axis. In addition to progressively changing optoelectronic properties, detected here by locally excited optical spectroscopy, our gated transport results for the GeS wires demonstrate that van der Waals nanowires readily support measurements of charge transport along helical twist moiré patterns (Extended Data Fig. 9). Hence, chiral nanowires represent a versatile platform for exploring phenomena associated with variable interlayer twist in layered materials.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-019-1147-x>.

Received: 21 November 2018; Accepted: 25 January 2019;

Published online 22 April 2019.

- Dean, C. R. et al. Hofstadter's butterfly and the fractal quantum Hall effect in moiré superlattices. *Nature* **497**, 598–602 (2013).
- Liu, K. et al. Evolution of interlayer coupling in twisted molybdenum disulfide bilayers. *Nat. Commun.* **5**, 4966 (2014).
- Yeh, P.-C. et al. Direct measurement of the tunable electronic structure of bilayer MoS_2 by interlayer twist. *Nano Lett.* **16**, 953–959 (2016).
- Carr, S. et al. Twistrionics: manipulating the electronic properties of two-dimensional layered structures through their twist angle. *Phys. Rev. B* **95**, 075420 (2017).
- Cao, Y. et al. Unconventional superconductivity in magic-angle graphene superlattices. *Nature* **556**, 43–50 (2018).
- Cao, Y. et al. Correlated insulator behaviour at half-filling in magic-angle graphene superlattices. *Nature* **556**, 80–84 (2018).
- Ribeiro-Palau, R. et al. Twistable electronics with dynamically rotatable heterostructures. *Science* **361**, 690–693 (2018).
- Yankowitz, M. et al. Emergence of superlattice Dirac points in graphene on hexagonal boron nitride. *Nat. Phys.* **8**, 382–386 (2012).
- Ponomarenko, L. A. et al. Cloning of Dirac fermions in graphene superlattices. *Nature* **497**, 594–597 (2013).
- Kim, K. et al. van der Waals heterostructures with high accuracy rotational alignment. *Nano Lett.* **16**, 1989–1995 (2016).
- Kang, K. et al. Layer-by-layer assembly of two-dimensional materials into wafer-scale heterostructures. *Nature* **550**, 229–233 (2017).

12. Frisenda, R. et al. Recent progress in the assembly of nanodevices and van der Waals heterostructures by deterministic placement of 2D materials. *Chem. Soc. Rev.* **47**, 53–68 (2018).
13. Sutter, E. & Sutter, P. 1D wires of 2D layered materials: germanium sulfide nanowires as efficient light emitters. *ACS Appl. Nano Mater.* **1**, 1042–1049 (2018).
14. Eshelby, J. D. Screw dislocations in thin rods. *J. Appl. Phys.* **24**, 176–179 (1953).
15. Bierman, M. J., Lau, Y. K. A., Kvit, A. V., Schmitt, A. L. & Jin, S. Dislocation-driven nanowire growth and Eshelby twist. *Science* **320**, 1060–1063 (2008).
16. Wagner, R. S. & Ellis, W. C. Vapor–liquid–solid mechanism of single crystal growth. *Appl. Phys. Lett.* **4**, 89–90 (1964).
17. Morales, A. M. & Lieber, C. M. A laser ablation method for the synthesis of crystalline semiconductor nanowires. *Science* **279**, 208–211 (1998).
18. Suryawanshi, S. R., Warule, S. S., Patil, S. S., Patil, K. R. & More, M. A. Vapor–liquid–solid growth of one-dimensional tin sulfide (SnS) nanostructures with promising field emission behavior. *ACS Appl. Mater. Interfaces* **6**, 2018–2025 (2014).
19. Liu, S. et al. Solution-phase synthesis and characterization of single-crystalline SnSe nanowires. *Angew. Chem. Int. Ed.* **50**, 12050–12053 (2011).
20. Peng, H., Meister, S., Chan, C. K., Zhang, X. F. & Cui, Y. Morphology control of layer-structured gallium selenide nanowires. *Nano Lett.* **7**, 199–203 (2007).
21. Zhai, T. et al. Fabrication of high-quality In₂Se₃ nanowire arrays toward high-performance visible-light photodetectors. *ACS Nano* **4**, 1596–1602 (2010).
22. Yang, R. B. et al. Pulsed vapor–liquid–solid growth of antimony selenide and antimony sulfide nanowires. *Adv. Mater.* **21**, 3170–3174 (2009).
23. Alegria, L. D., Yao, N. & Petta, J. R. MOCVD synthesis of compositionally tuned topological insulator nanowires. *Phys. Status Solidi B*, 991–996 (2014).
24. Burton, W. K., Cabrera, N. & Frank, F. C. The growth of crystals and the equilibrium structure of their surfaces. *Phil. Trans. R. Soc. Lond. A* **243**, 299–358 (1951).
25. Bletskan, D. Production of GeS single crystals, investigation of their morphology and of latter influence on hologram recording. *Kristallografiya* **20**, 1008–1012 (1975).
26. Sutter, P. & Sutter, E. Growth mechanisms of anisotropic layered group IV chalcogenides on van der Waals substrates for energy conversion applications. *ACS Appl. Nano Mater.* **1**, 3026–3034 (2018).
27. Jain, A. et al. Commentary. The Materials Project: a materials genome approach to accelerating materials innovation. *APL Mater.* **1**, 011002 (2013).
28. Bissert, G. & Hesse, K. F. Verfeinerung der struktur von germanium(II)-sulfid, GeS. *Acta Crystallogr. B* **34**, 1322–1323 (1978).
29. Sutter, P., Argyropoulos, C. & Sutter, E. Germanium sulfide nano-optics probed by STEM-cathodoluminescence spectroscopy. *Nano Lett.* **18**, 4576–4583 (2018).
30. Kociak, M. & Zagonel, L. F. Cathodoluminescence in the scanning transmission electron microscope. *Ultramicroscopy* **176**, 112–131 (2017).
31. Hentschel, M., Schäferling, M., Duan, X., Giessen, H. & Liu, N. Chiral plasmonics. *Sci. Adv.* **3**, e1602735 (2017).
32. Sutter, E. A. & Sutter, P. W. Size-dependent phase diagram of nanoscale alloy drops used in vapor–liquid–solid growth of semiconductor nanowires. *ACS Nano* **4**, 4943–4947 (2010).

Acknowledgements The synthesis of twisted GeS nanowires, analysis of nanobeam diffraction and transport measurements were supported by the National Science Foundation, Division of Materials Research, Solid State and Materials Chemistry Program under grant no. DMR-1607795. The development of nanobeam electron diffraction correlated with locally excited CL spectroscopy measurements and of associated data analysis methods was supported by the US Department of Energy, Office of Science, Basic Energy Sciences, under award no. DE-SC0016343.

Reviewer information Nature thanks Hua Zhang and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions P.S. and E.S. planned the study, carried out the nanowire growth and the measurements, and analysed the data. S.W. analysed the electron diffraction data and performed device fabrication and transport measurements. P.S. and E.S. wrote the paper. All authors commented on the manuscript.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-019-1147-x>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to P.S. or E.S. **Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

METHODS

Synthesis. GeS nanowires were synthesized by vapour–liquid–solid growth through sublimation of GeS powder (99.99%, Sigma-Aldrich) in a pumped quartz tube reactor with two temperature zones. In zone 1, a quartz boat with GeS powder (roughly 50 mg) was heated to 450 °C. Zone 2, containing the substrate, was heated to temperatures between 270 °C and 350 °C. Si(100) wafers covered with 2–4-nm-thick Au films, deposited by sputtering at room temperature and dewetted at the growth temperature, were used as substrates. During growth, a carrier gas flow (Ar, 2% H₂) of 50 standard cubic centimetres per minute was maintained at a pressure of 20 mTorr. Typical growth times of 10 min produced forests of nanowires with lengths of several tens of micrometres. Additional details on growth and characterization can be found elsewhere¹³. The as-grown nanowires contained ubiquitous axial screw dislocations. Among more than 100 GeS nanowires considered, all were found to contain such dislocations. Nanowires for analysis by diffraction and cathodoluminescence were randomly chosen to cover a range of wire diameters and lengths.

Electron microscopy and diffraction. The morphology of the nanowires was investigated by (S)TEM (FEI Talos F200X). Series of nanobeam diffraction patterns were obtained using an incident electron beam with size <3 nm, displaced in predefined equal steps along the centre axis of individual nanowires. Nanobeam diffraction patterns were analysed using the software package JEMS. First, the nearest apparent zone axis of each pattern was determined. The simulated sample was then tilted to match the experimental diffraction pattern and determine the actual zone axis. To calculate the twist, each zone axis $[h\ k\ l]$ was considered as $[h\ k\ 0]$, and the angle to the $[1\ 0\ 0]$ axis was determined by the dot product.

Cathodoluminescence spectroscopy. Cathodoluminescence (CL) spectroscopy was performed in STEM mode (STEM-CL) using a Gatan Vulcan CL holder between $T = 110\text{ K}$ and 300 K at 200 keV electron energy. The incident beam current for CL measurements was typically 300–400 pA. Spectrum line scans were acquired by displacing the electron beam in predefined equal steps along individual nanowires and acquiring full CL spectra at each beam position. Typical acquisition times were 10 s per spectrum. Apart from changes in signal-to-noise ratio, no changes to spectral characteristics were observed for longer or shorter integration times.

Delocalization of charge carriers in STEM-CL line scans. To examine the origin of the broad luminescence peaks observed in STEM-CL on twisted GeS nanowires, control experiments were performed on a long (about 5 μm) GeS nanowire (Extended Data Fig. 7a). Rather than scanning the exciting electron beam continuously from one end of the wire to the other, five individual STEM-CL spectrum line scans were measured as shown in Extended Data Fig. 7b, that is, with top-to-bottom scan direction and stitching line scans 1–5 together from bottom to top. This scan pattern was chosen to identify possible displacements of carrier recombination and light emission from the position of the exciting electron beam.

The composite of the five spectrum line scans shown in Extended Data Fig. 7b shows repeated spectral features in consecutive scans, even though the electron-beam excitation was seamlessly joined without any overlap. This repetition indicates a spreading of the minority carriers (electrons in the p-type GeS nanowires, as shown by gated transport measurements, Extended Data Fig. 9). Instead of a symmetric bidirectional spreading as expected for delocalization by electron diffusion, Extended Data Fig. 7b shows an asymmetric displacement of the spectral features. This suggests an asymmetric spreading of locally generated carriers, either due to drift in an electric field along the nanowire axis or as a result of a spatially varying diffusion coefficient D (Extended Data Fig. 10). Both could be caused by the filling of trap states along the nanowire (for example at the core of the axial

dislocation) by electron-beam excited carriers. An electric field would drive carrier drift along the wire. A transient reduction in carrier mobility (μ) due to scattering by charged traps³³ would, via the Einstein relationship, $\frac{D}{\mu} = \frac{kT}{q}$, translate to a spatially varying diffusion coefficient (D), causing preferential minority carrier diffusion away from regions excited previously by the electron beam. Additional work is needed to clarify the origin of the directional spreading and recombination of electron-beam excited charge carriers along twisted van der Waals nanowires.

Extended Data Fig. 7c shows a lowest-order ‘corrected’ composite STEM-CL spectrum line scan in which the repeated spectral features have been removed by cutting at dashed lines in Extended Data Fig. 7b, and the overall length re-adjusted to match the dimensions of the nanowire.

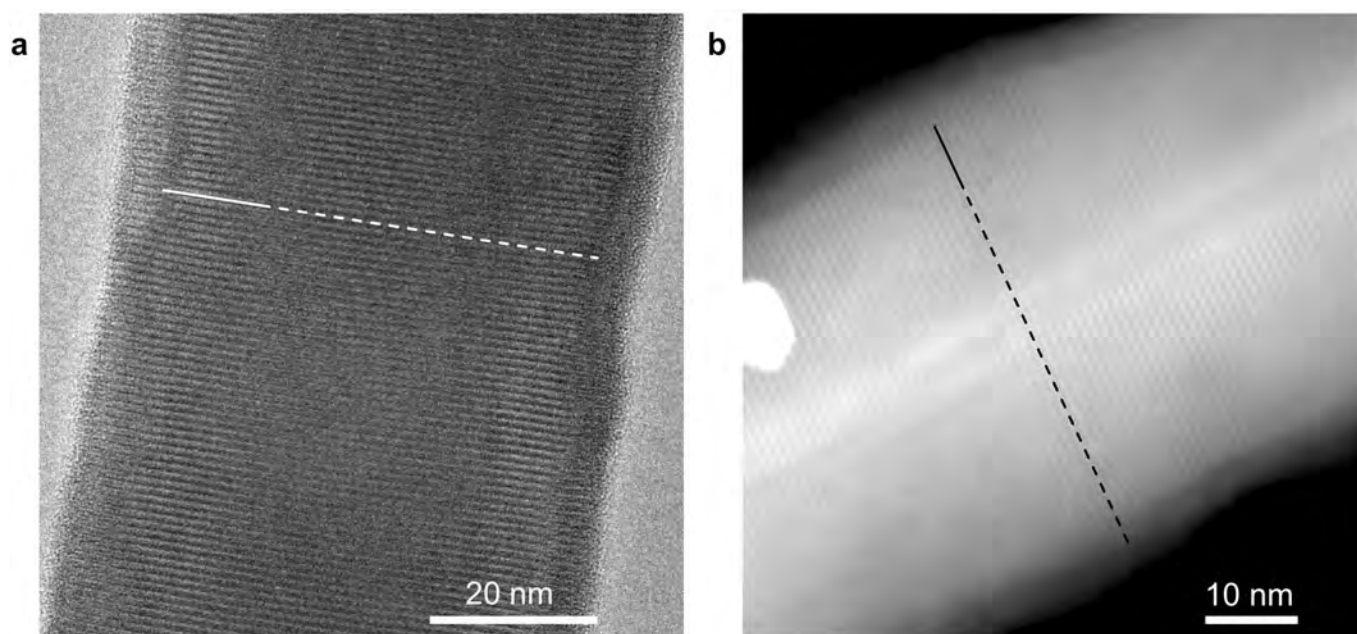
Correlation of optoelectronic properties with twist moiré registry. In addition to the data shown in Fig. 3 and Extended Data Fig. 3, series of CL spectra obtained on long GeS nanowires were used to identify signatures of varying electronic structure due to the twist moiré realized in chiral van der Waals nanowires. An example of this type of analysis is shown in Extended Data Fig. 8. Extended Data Fig. 8a shows a diffraction analysis of the cumulative twist along a long GeS nanowire segment. Over a length of 4.6 μm , a cumulative twist of 155° is observed. Extended Data Fig. 8b is the composite STEM-CL spectrum line scan along the wire, obtained as shown in Extended Data Fig. 7. The Burgers vector analysis for this wire shows a pronounced peak at 1.1 nm, consistent with a Burgers vector equal to the GeS c -axis unit cell size (Extended Data Fig. 8c), that is, a helical structure with pitch 2L (two GeS monolayers).

In contrast to planar twisted van der Waals heterostructures, in which interlayer twist gives rise to a 2D moiré pattern, chiral layered nanowires project the twist moiré onto a helical van der Waals interface. The path across the moiré followed by the centre of the helix corresponds to the arc in Extended Data Fig. 8d, while the periphery of the helix sweeps along a spiral, as illustrated in Extended Data Fig. 8e. The moiré pattern and arc shown in Extended Data Fig. 8d are drawn to scale for the long chiral GeS nanowire of Extended Data Fig. 8a. Evidently, between points with zone axes $[010]$ and $[100]$, that is, a segment of about 2.6 μm length, the nanowire sweeps through several regions with closely aligned and misaligned GeS lattices at the van der Waals interface (see Extended Data Fig. 8f). The positions of the closely aligned regions of the twist moiré, marked with arrows on the STEM-CL line scan of Extended Data Fig. 8b, show a close correspondence with the luminescence intensity maxima along the nanowire. Other spectral features such as the centre wavelength or FWHM are more sensitive to competing effects, including, for instance, a progressive change in diameter of the slightly tapered nanowire causing an overall redshift of the spectral centre, and local differences in the spreading of excited charge carriers causing differences in peak width. Note that near the end of the nanowire (as in Fig. 3, for example), the carrier delocalization is limited so that the electronic effects of a changing moiré registry are visible in the peak centre, intensity and FWHM.

Data availability

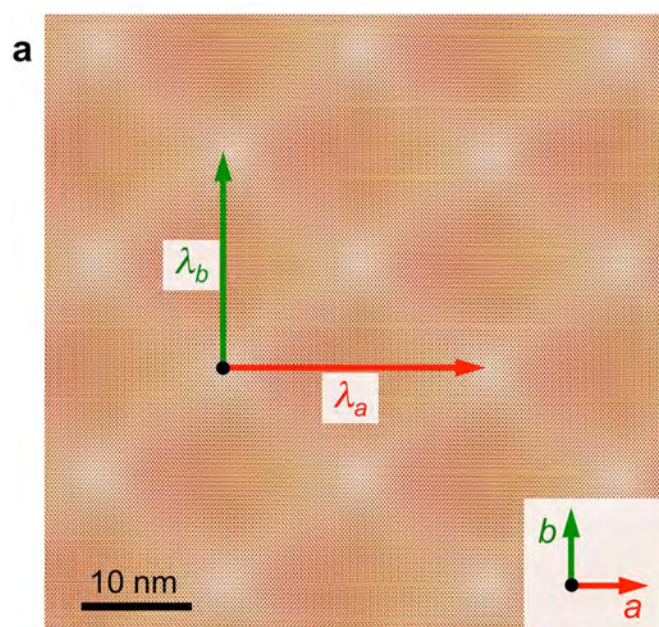
Raw data are included for Figs. 1 and 3a, b, and Extended Data Figs. 1, 3, 4a, b, 5, 6a–e, 7 and 9. The full set of diffraction data that support the nanowire twist analysis is available from the corresponding author upon request.

33. Wei, P.-C. et al. Room-temperature negative photoconductivity in degenerate InN thin films with a supergap excitation. *Phys. Rev. B* **81**, 045306 (2010).
34. Tan, D. et al. Anisotropic optical and electronic properties of two-dimensional layered germanium sulfide. *Nano Res.* **10**, 546–555 (2017).

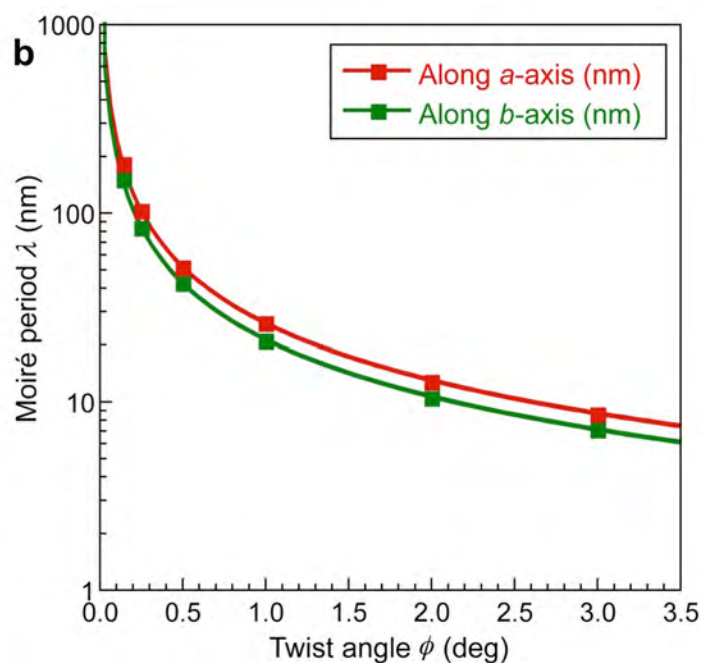


Extended Data Fig. 1 | Electron microscopy of twisted GeS nanowires with axial screw dislocations. **a**, High-resolution TEM image of a 65-nm GeS nanowire. The lattice fringes along the wire axis are spaced by 1.06 nm

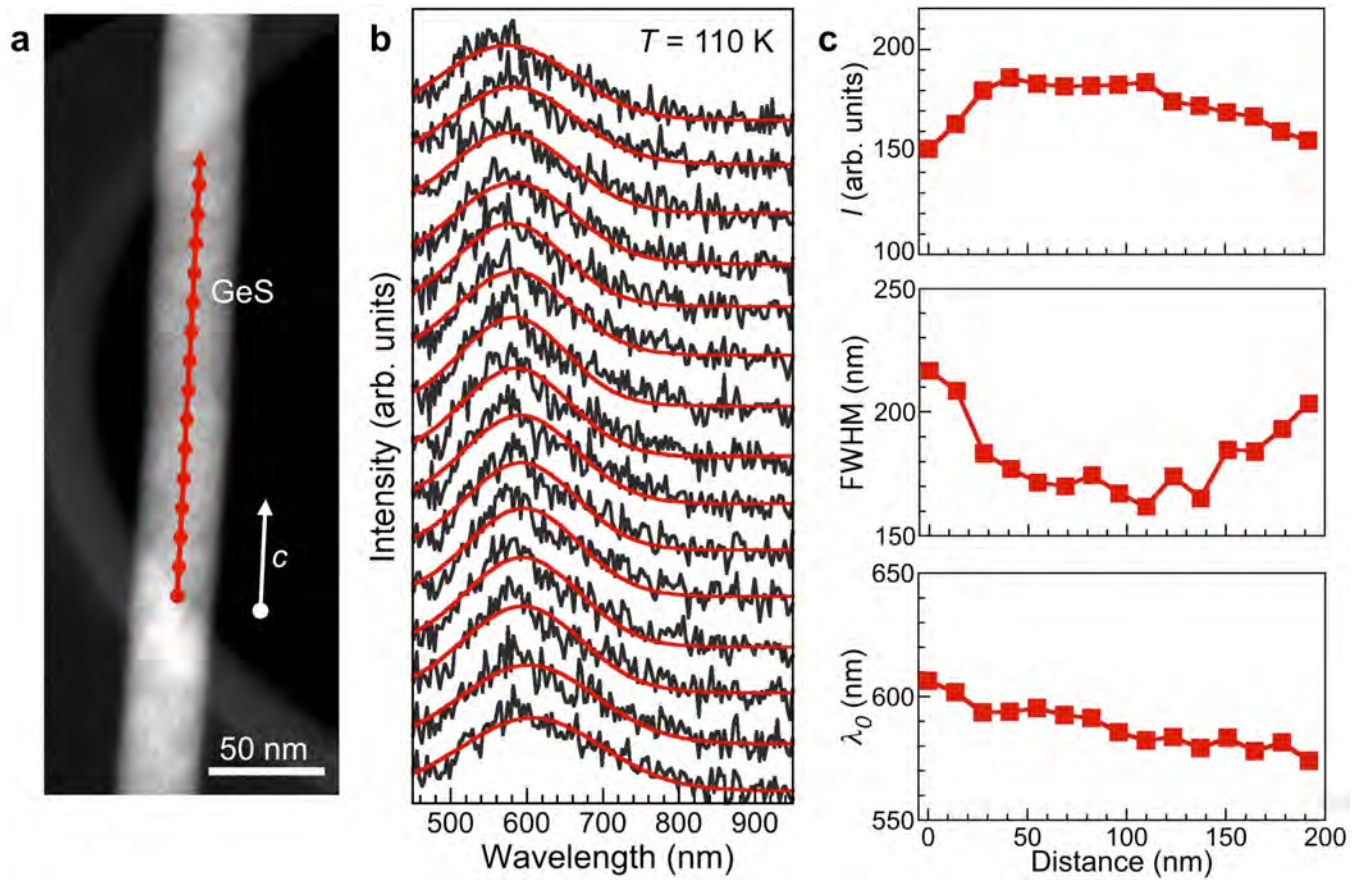
(that is, the GeS *c*-axis unit cell dimension). **b**, High-resolution STEM image of a 54-nm GeS nanowire. As in **a**, the lattice fringes along the wire axis are spaced by the GeS *c*-axis unit cell dimension.



Extended Data Fig. 2 | GeS twist moiré patterns. **a**, Illustration of a GeS twist moiré (here with 1° twist angle). Note the reversal of the ratio $\lambda_a/\lambda_b = b/a$, where (a, b) denote the in-plane lattice parameters of GeS, and λ_a, λ_b are the moiré periods along the a and b directions. **b**, Moiré

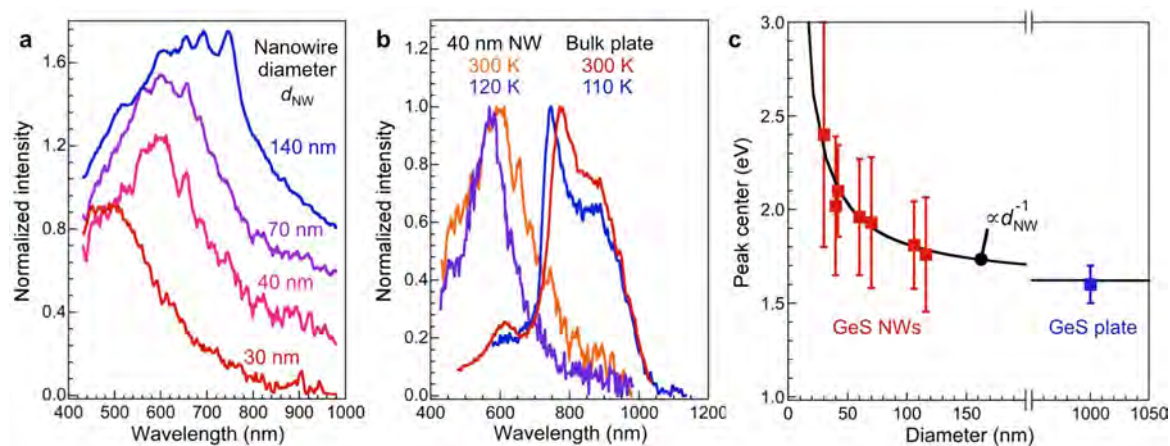


periodicity for a twisted GeS van der Waals interface as a function of interlayer twist angle. Data points were obtained from overlays of models of two GeS monolayers with different relative orientation. The solid line is a power-law fit $\lambda \propto \phi^{-1}$ to the data.



Extended Data Fig. 3 | STEM-CL line scan along a twisted GeS nanowire measured at $T = 110$ K. **a**, HAADF-STEM image of a GeS nanowire (diameter about 44 nm). **b**, Position-dependent STEM-CL spectra recorded along the nanowire at points marked in **a** ($T = 110$ K). Lines are

Gaussian fits to the dominant band-edge luminescence peak. **c**, Plots of the Gaussian fit parameters (amplitude I ; FWHM; centre wavelength λ_0) as a function of position along the nanowire.

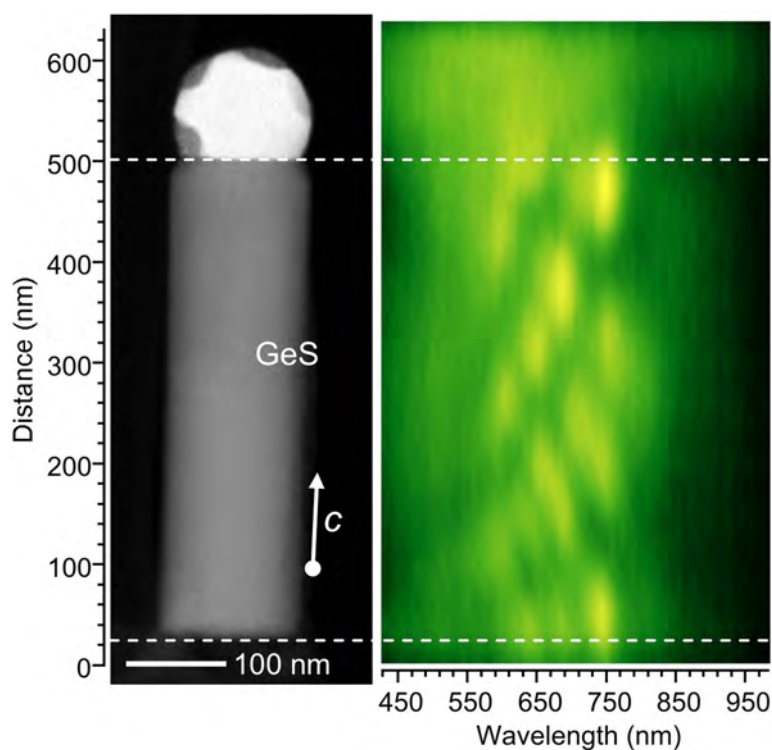


Extended Data Fig. 4 | Size-dependent GeS band-edge luminescence.

a, Examples of STEM-CL spectra for GeS nanowires with four different diameters (140 nm, 70 nm, 40 nm, 30 nm) measured at room temperature. Note the progressive blueshift with decreasing nanowire diameter and the intensity modulation for the 140-nm wire, which is due to photonic waveguide mode interference in thick nanowires similar to that observed in planar GeS plates and prisms²⁹ (see Extended Data Fig. 5).

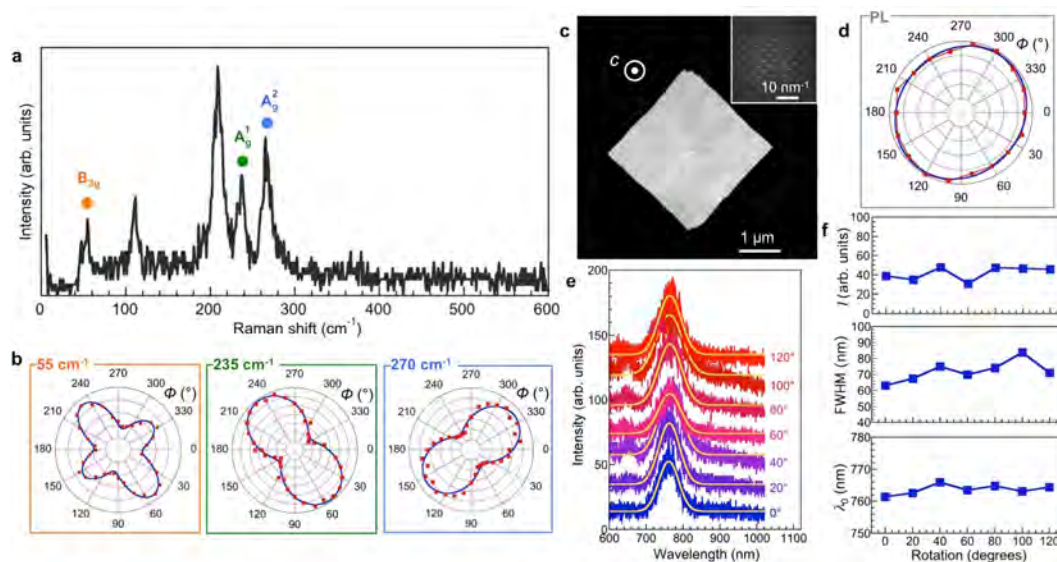
b, Comparison of the temperature-dependent CL spectra of a 40-nm-diameter GeS nanowire and of a micrometre-sized, single-crystalline

bulk GeS plate; CL spectra were measured at room temperature and at $T = 120$ K and 110 K. For nanowires and plates, the luminescence peak narrows and blueshifts by the same amount (about 30 nm) at low temperature. **c**, Analysis of the peak photon energy of the CL spectra of GeS nanowires with different diameters, d_{NW} and of a bulk GeS plate (lateral size approximately 1 μm). Error bars are based on the FWHM of the luminescence peaks. The black line is a fit $h\nu \propto d_{\text{NW}}^{-1}$, with a bulk bandgap of 1.60 eV (ref. ²⁹).



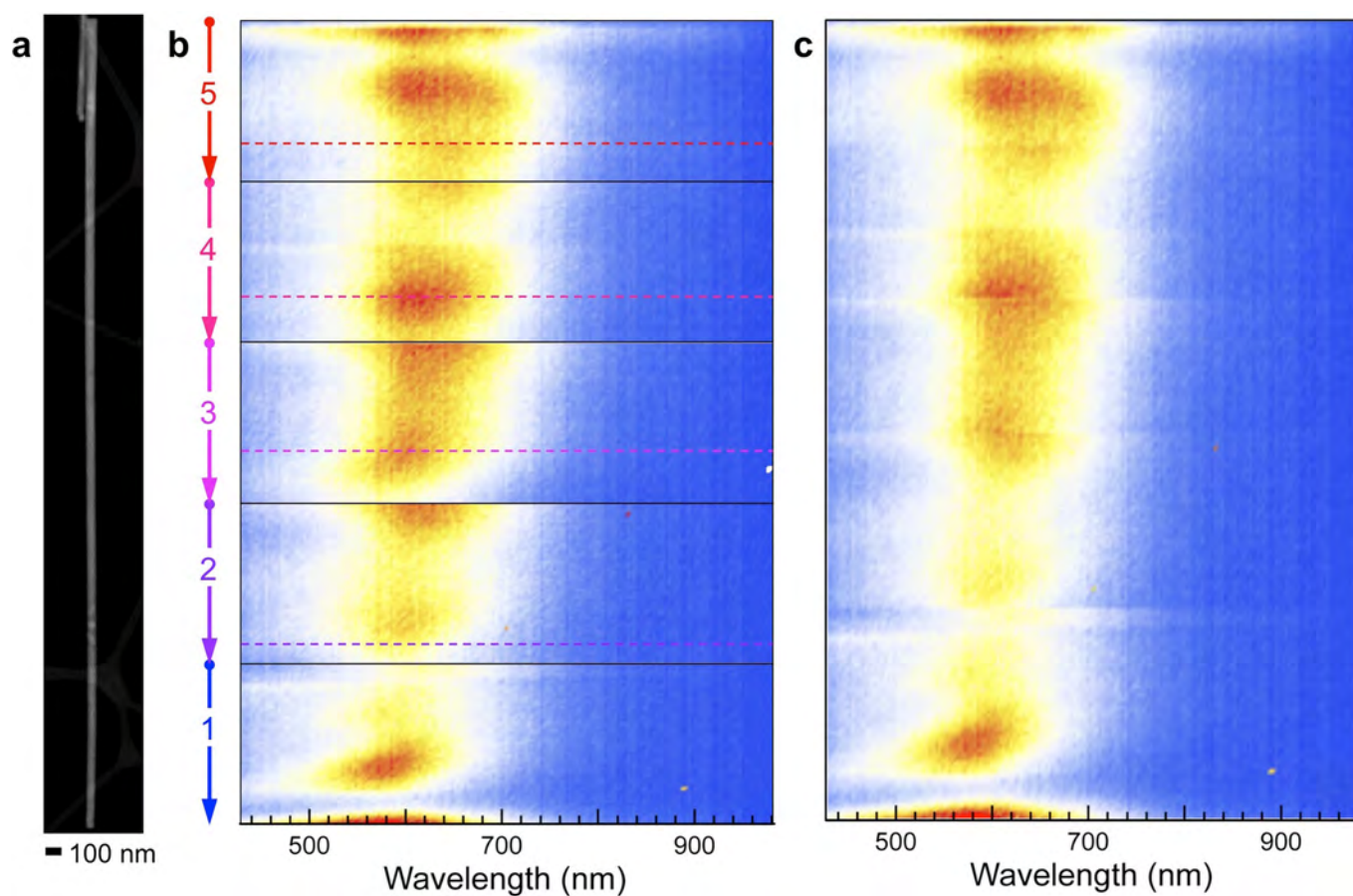
Extended Data Fig. 5 | Waveguide-mode interference in thick GeS nanowires. HAADF-STEM image of a 140-nm-thick nanowire near the Au-rich VLS catalyst (left), and STEM-CL line scan along the centre axis of

the wire (right) showing fringes due to interference of travelling waveguide modes reflected by the specular end facets of the nanowire²⁹.



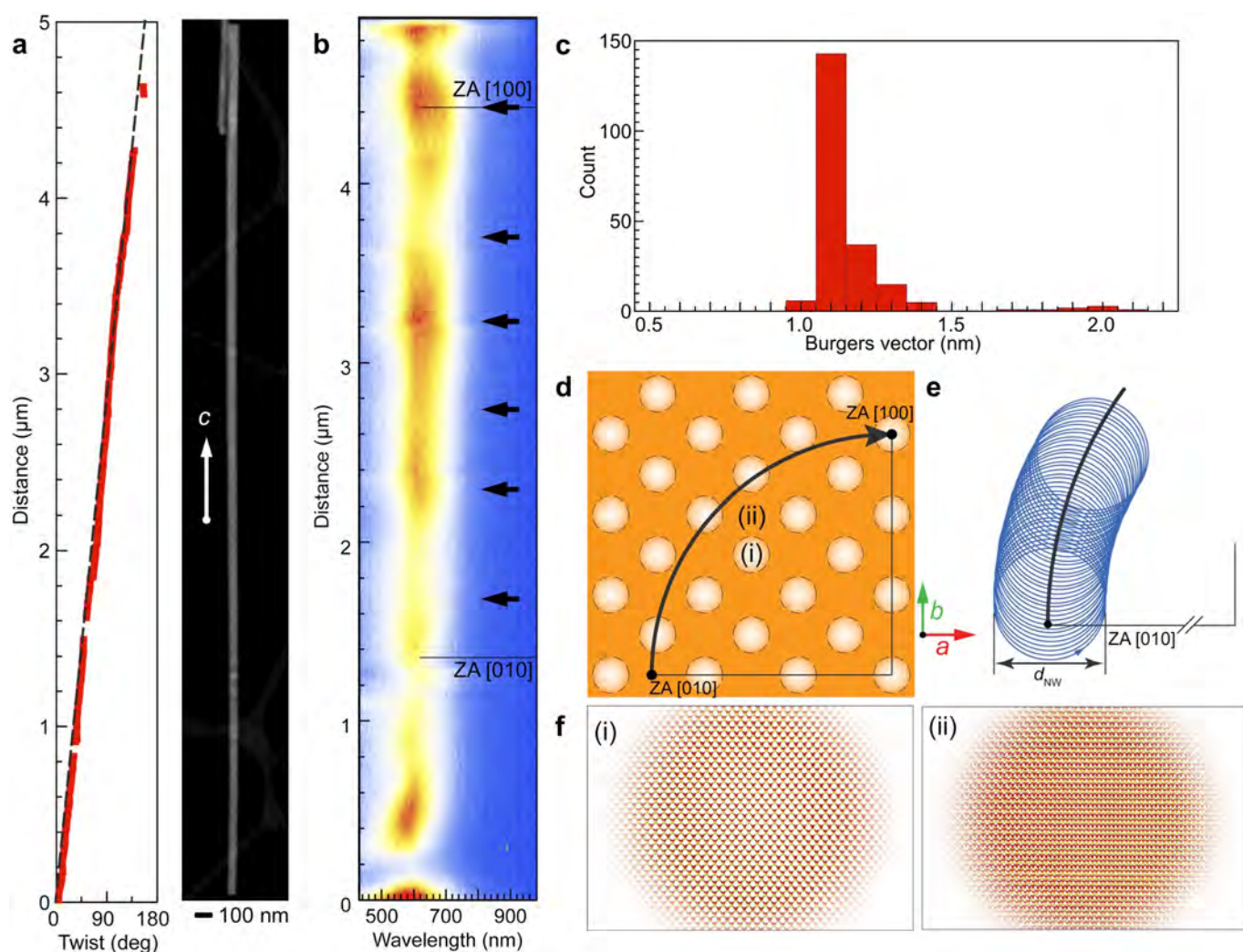
Extended Data Fig. 6 | Polarized Raman and photoluminescence on mesoscale monocrystalline GeS plates. **a**, Raman spectrum on a monocrystalline GeS plate with linearly polarized laser light ($\lambda = 532$ nm) incident along the c axis, showing the B_{2g} and A_g modes associated with the orthorhombic layered crystal. **b**, Polar plots of Raman intensity as a function of the polarization direction of the incident light for three different modes. Note the strongly anisotropic Raman scattering by the single-crystalline GeS plate³⁴. **c**, HAADF-STEM image of a mesoscale GeS plate. The c axis lies perpendicular to the large top facet of the plate. Inset: electron diffraction pattern of the GeS plate. **d**, Polar plot of band-

edge photoluminescence intensity as a function of the polarization of the exciting laser beam ($\lambda = 532$ nm). **e**, Photoluminescence spectra for different orientation of the GeS plate relative to the polarization of the incident light. Lines show Gaussian fits to the main peak. **f**, Plots of the parameters of the Gaussian fits for different rotation angles of the plate relative to the polarization of the incident light: amplitude (I), FWHM, centre wavelength (λ_0). Polarization dependences were measured on individual GeS plates using a Horiba Scientific XPlora Plus Raman/photoluminescence microscope with angle between sample and incident light polarization varied using a sample rotation stage.



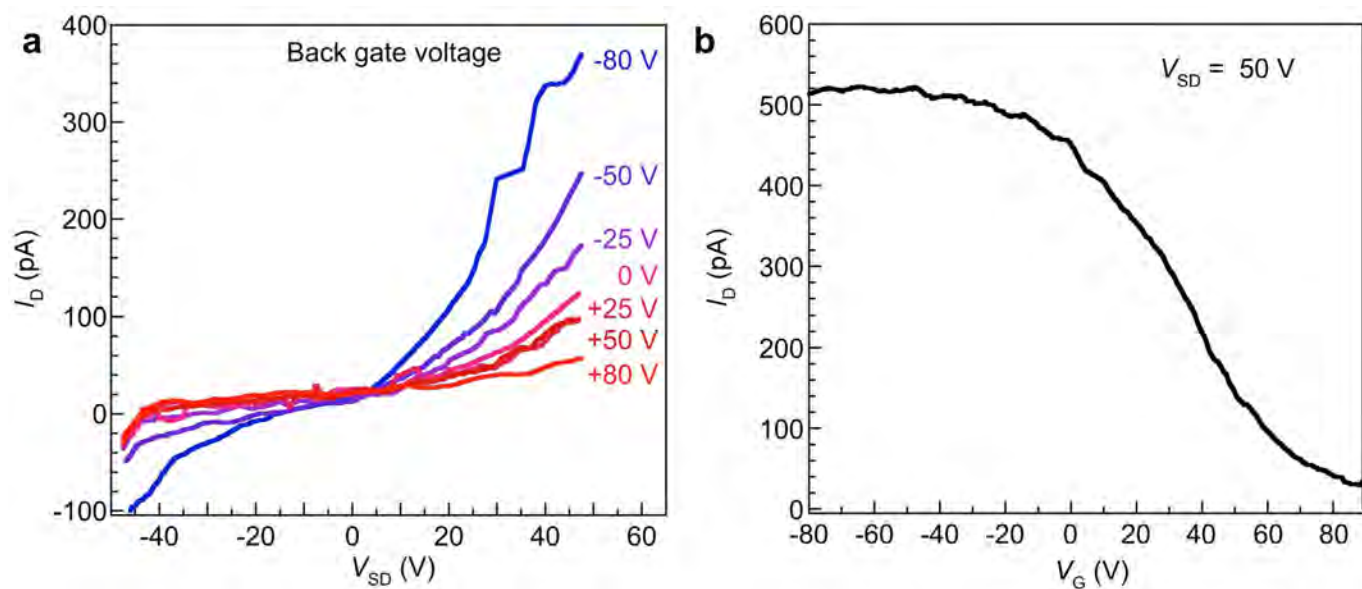
Extended Data Fig. 7 | STEM-CL line scans on a long GeS nanowire ($T = 300$ K). **a**, HAADF-STEM image of the GeS nanowire (approximately $5\ \mu\text{m}$ long). **b**, STEM-CL spectrum line scan, obtained in five consecutive sections scanned as shown by arrows. The particular scan pattern was chosen to identify displacements of charge-carrier recombination from

the position of the exciting electron beam. Note the repetition of spectral features, even though the electron-beam excitation was precisely stitched in consecutive line scans. **c**, Reconstruction of the CL spectra along the nanowire by cutting repeated sections along dashed lines shown in **b**.



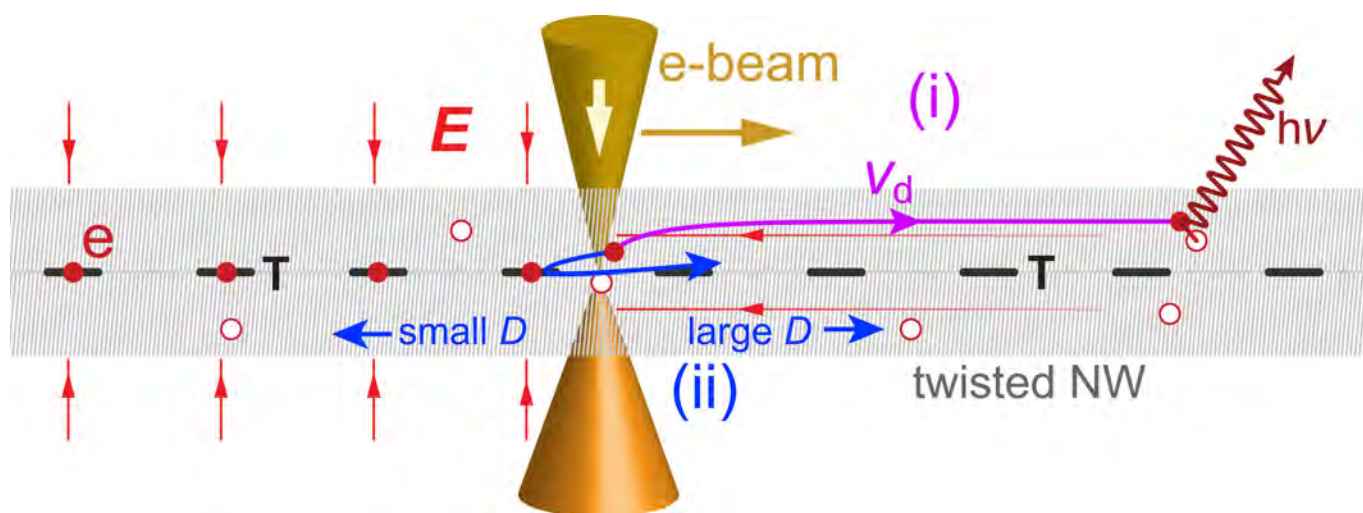
Extended Data Fig. 8 | Twist and STEM-CL analysis of a long GeS nanowire section (length about 5 μm , $T = 300$ K). **a**, HAADF-STEM image of the GeS nanowire and analysis of the cumulative twist angle along the wire. **b**, STEM-CL spectrum line scan of the twisted GeS nanowire, reconstructed from five individual line scans as shown in Extended Data Fig. 7. The twist angle increases by 90° between points with zone axis [010] and [100]. **c**, Histogram of Burgers vectors along the GeS nanowire shown in **a**, using the same analysis as in Fig. 2. **d**, Evolution of the twist moiré registry along the nanowire. White and orange shaded regions indicate sections of a planar twist moiré pattern in which the atoms in the adjacent

GeS layers are closely aligned (white, (i)) or misaligned (orange, (ii)). The arc indicates the progression of the moiré registry between zone axis [010] and [100] along the nanowire; it sweeps through several regions with aligned and misaligned GeS lattices. The positions of closely aligned regions, marked by arrows in **b**, coincide approximately with intensity maxima of the STEM-CL spectrum line scan. **e**, Illustration of parts of the spiral path of the periphery of the helical van der Waals interface across the twist moiré pattern (d_{NW} , nanowire diameter). **f**, Illustration of the twist moiré pattern. (i) shows the closely aligned regions, and (ii) shows the misaligned regions of the twist moiré (drawn here for 0.25° interlayer twist).



Extended Data Fig. 9 | Gated charge transport in twisted GeS nanowires. **a**, Drain current as a function of source-drain bias measured at room temperature for a single GeS nanowire field-effect transistor (FET) back-gated via 300 nm SiO_2/Si . Nanowire diameter, 116 nm. Channel length, 5 μm . **b**, Transfer characteristics (drain current as a function of

back-gate voltage) for the GeS nanowire FET. For transport measurements, Ti/Au contacts (10 nm/60 nm) were deposited on single GeS nanowires using patterning with electron-beam lithography; transport measurements were made on a Lakeshore probe station under vacuum (1×10^{-7} Torr) at room temperature.



Extended Data Fig. 10 | Schematic of asymmetric minority carrier drift or diffusion along the GeS nanowires. Illustration of the two possible mechanisms discussed in the Methods, both related to the filling of charge traps (T) by electron-beam excited carriers: (i) drift (v_d) due to a net electric field along the wire axis; or (ii) charged impurity scattering causing

differences in carrier mobility and diffusion coefficient (D). In line scans as shown in Extended Data Fig. 7, both mechanisms would cause light emission predominantly downstream (that is, in areas not yet scanned by the electron beam) from the excitation spot.

Helical van der Waals crystals with discretized Eshelby twist

Yin Liu^{1,2,9}, Jie Wang^{3,9}, Sujung Kim^{1,8,9}, Haoye Sun^{1,9}, Fuyi Yang^{1,2}, Zixuan Fang^{1,4}, Nobumichi Tamura⁵, Ruopeng Zhang^{1,6}, Xiaohui Song⁶, Jianguo Wen³, Bo Z. Xu¹, Michael Wang¹, Shuren Lin^{1,2}, Qin Yu², Kyle B. Tom^{1,2}, Yang Deng¹, John Turner⁶, Emory Chan⁷, Dafei Jin³, Robert O. Ritchie^{1,2}, Andrew M. Minor^{1,6}, Daryl C. Chrzan^{1,2}, Mary C. Scott^{1,6} & Jie Yao^{1,2*}

The ability to manipulate the twisting topology of van der Waals structures offers a new degree of freedom through which to tailor their electrical and optical properties. The twist angle strongly affects the electronic states, excitons and phonons of the twisted structures through interlayer coupling, giving rise to exotic optical, electric and spintronic behaviours^{1–5}. In twisted bilayer graphene, at certain twist angles, long-range periodicity associated with moiré patterns introduces flat electronic bands and highly localized electronic states, resulting in Mott insulating behaviour and superconductivity^{3,4}. Theoretical studies suggest that these twist-induced phenomena are common to layered materials such as transition-metal dichalcogenides and black phosphorus^{6,7}. Twisted van der Waals structures are usually created using a transfer-stacking method, but this method cannot be used for materials with relatively strong interlayer binding. Facile bottom-up growth methods could provide an alternative means to create twisted van der Waals structures. Here we demonstrate that the Eshelby twist, which is associated with a screw dislocation (a chiral topological defect), can drive the formation of such structures on scales ranging from the nanoscale to the mesoscale. In the synthesis, axial screw dislocations are first introduced into nanowires growing along the stacking direction, yielding van der Waals nanostructures with continuous twisting in which the total twist rates are defined by the radii of the nanowires. Further radial growth of those twisted nanowires that are attached to the substrate leads to an increase in elastic energy, as the total twist rate is fixed by the substrate. The stored elastic energy can be reduced by accommodating the fixed twist rate in a series of discrete jumps. This yields mesoscale twisting structures consisting of a helical assembly of nanoplates demarcated by atomically sharp interfaces with a range of twist angles. We further show that the twisting topology can be tailored by controlling the radial size of the structure.

Many crystals can be grown into twisted forms on scales ranging from nanoscale to mesoscale and macroscale^{8–14}. Specifically, an axial screw dislocation in a one-dimensional structure can produce a continuous crystallographic twist, known as the Eshelby twist, giving rise to helically twisted nanowire structures^{15,16}. This mechanism potentially provides a means to create twisted van der Waals (vdW) structures. Materials such as germanium sulfide (GeS) can grow into nanowires along the vdW stacking direction (the cross-plane direction)^{17–20}, and introducing Eshelby twist into such nanowires naturally leads to twist between the successive layers. Even though screw dislocations in layered vdW materials are well known^{21–23}, the Eshelby twist has not been considered in those materials. Here we report the observation of the Eshelby twist in vdW materials and show how this enables the synthesis of vdW structures with various twisting morphologies.

This approach opens new possibilities for creating twisted vdW structures with tailored topologies.

The synthesis method was demonstrated using GeS, a layered IV–VI monochalcogenide. Twisted GeS structures were synthesized on silicon substrates with a 3–4 nm native oxidized layer by using a chemical vapour transport method with gold as a catalyst (see Methods and Extended Data Fig. 1). A representative scanning electron microscopy (SEM) image of a mesoscale crystal is presented in Fig. 1a, showing a well-defined helicoidal morphology. The synthesized crystals have varying twist periods ranging from 2 μm to 20 μm , with total lengths up to hundreds of micrometres, and radial sizes ranging from several hundred nanometres to more than 10 μm . They have a three-dimensional architecture consisting of periodically rotating nanoplates with a thickness of several hundred nanometres, as is revealed by cross-sectional transmission electron microscopy (TEM) (Fig. 1b) and cross-sectional SEM images acquired through consecutive focused ion beam milling along the twist axis (Fig. 1d, Supplementary Video 1). Statistical analysis suggests approximately equal numbers of left-handed and right-handed crystals (Extended Data Fig. 2). Quantitative chemical analysis of the structure indicates a 1:1 atomic ratio of Ge:S (Extended Data Fig. 2). The crystallography is revealed through synchrotron X-ray Laue microdiffraction analysis with submicrometre spatial resolution²⁴. X-ray crystal orientation maps of the structures (Fig. 1e, f, Extended Data Fig. 2) show that the twist axis of the crystal is along the *c* axis (the cross-plane direction), and the vdW planes, defined by the *a* and *b* axes in Fig. 1c, periodically rotate about the *c* axis such that there is a total twist of 180° in a single period (that is, between two adjacent minimum widths as seen in Fig. 1a).

Atomic-resolution scanning transmission electron microscopy (STEM) suggests that the nanoplates are single crystals in the space group *Pcmn* and that the twist interface between the nanoplates is atomically sharp (Fig. 2a, Extended Data Fig. 2). The twist angle at this interface is 10.27° (as determined from the change of tilt angle needed to tilt each of the two crystals to the [010] zone axis; Fig. 2a, b). The twist interface was also characterized by plan-view TEM of the stacking nanoplates with the incident electron beam along the twist axis (Fig. 2c–e, Extended Data Fig. 3). Electron diffraction patterns indicate a misorientation angle of 7.5° between the two nanoplates (Fig. 2c, d), and double diffraction patterns were clearly observed from the twist boundary (Fig. 2e). The double diffraction in reciprocal space reflects long-range ordering with a period of 2.26 nm in real space, which agrees well with the simulated rotational moiré pattern (Fig. 2f, g). In addition to TEM, we used electron backscattering diffraction to measure twist angles (Extended Data Fig. 4). In total, 15 twist angles in several structures were quantified, ranging from 6.8° to 16° with an average of 10.3°. Further estimates of the twist angles were based on

¹Department of Materials Science and Engineering, University of California Berkeley, Berkeley, CA, USA. ²Materials Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA.

³Center for Nanoscale Materials, Nanoscience and Technology Division, Argonne National Laboratory, Lemont, IL, USA. ⁴National Engineering Research Center of Electromagnetic Radiation Control Materials, University of Electronic Science and Technology of China, Chengdu, China. ⁵Advanced Light Source, Lawrence Berkeley National Laboratory, Berkeley, CA, USA. ⁶National Center for Electron Microscopy, Molecular Foundry, Lawrence Berkeley National Laboratory, Berkeley, CA, USA. ⁷Molecular Foundry, Lawrence Berkeley National Laboratory, Berkeley, CA, USA.

⁸Present address: Department of Electrical and Computer Engineering, University of California Santa Cruz, Santa Cruz, CA, USA. ⁹These authors contributed equally: Yin Liu, Jie Wang, Sujung Kim, Haoye Sun. *e-mail: yaojie@berkeley.edu

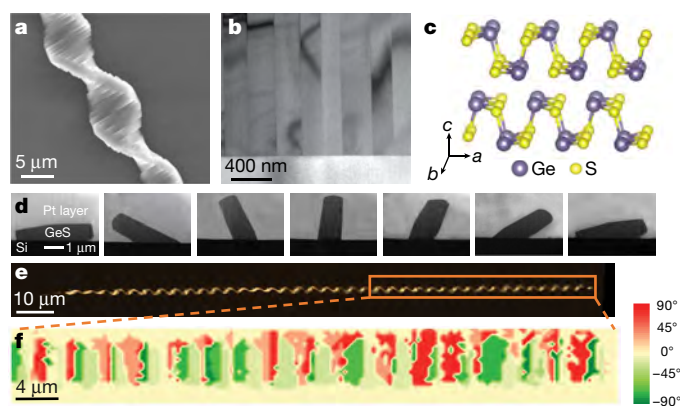


Fig. 1 | Structure of mesoscale twisted GeS crystals. **a**, SEM image showing the twisting morphology of a mesoscale GeS crystal. **b**, Cross-sectional TEM image of a twisted crystal. The surface normal of the cross-section is perpendicular to the twist axis. **c**, Layered crystal structure of GeS. **d**, A series of SEM images showing the evolution of radial cross-sections within one period of a twisted structure. **e**, **f**, Optical micrograph (**e**) and crystal orientation map (**f**) of a twisted structure generated by X-ray Laue microdiffraction analysis. The orientation angle is defined by the angle between the *b* axis and the norm of the substrate. The region of the X-ray analysis is highlighted by the orange box in **e**.

the number of nanoplates or interfaces per period, obtained from SEM imaging. The distribution of twist angles obtained this way (Fig. 2h) is in good agreement with the direct measurements.

Next, we show that this mesoscopic twist originates from Eshelby twist in the dislocated nanowires that were first grown via the gold-catalysed vapour–liquid–solid (VLS) method (Fig. 3a, b). In contrast to the VLS growth of GeS nanowires in previous studies^{17,18}, we observed GeS nanowires with continuous morphological twists (Fig. 3c), reminiscent of the Eshelby twist in non-vdW materials^{10,11}. Eshelby predicted that an axial screw dislocation at the centre of a thin whisker

will result in a twist that relieves the elastic energy associated with the screw dislocation¹⁶. The twist rate is $\alpha = \kappa b / A$ where *b* is the magnitude of the Burgers vector of the screw dislocation, κ is a prefactor related to the geometry of the whisker, and *A* is the cross-sectional area of the whisker. To grow nanowires with the Eshelby twist, we had to optimize growth parameters including pressures and carrier gas flow rates (see Methods); for non-optimal growth conditions, almost all of the synthesized nanowires are synthesized with no twist (Extended Data Fig. 5).

We confirmed the existence of the Eshelby twist through TEM studies. In TEM (Fig. 3d) and high-resolution TEM images (Fig. 3e), a line defect is evident at the centre of the nanowire. Convergent beam electron diffraction (CBED) shows that the nanowire has a crystallographic twist (Fig. 3d), giving rise to 24° of rotation about the *c* axis over an approximate distance of 1.4 μm. This amounts to a twist rate of 0.3 rad μm^{−1}, comparable to the twist rates reported in nanowires of conventional covalent semiconductors^{10,11}. Burgers vector analysis was performed on the basis of *g* · *b* contrast in TEM images (Fig. 3f–i): the dislocation is imaged with different *g* reflections in reciprocal space and becomes invisible when *g* · *b* = 0. Dark-field images were taken for *g* = (002) and *g* = (2̄10) for a strong two-beam condition near the [210] zone axis. At *g* = (002), high contrast of the dislocation was observed, whereas for *g* = (2̄10), the dislocation became invisible. As such, the direction of Burgers vector is determined to be along the [001] growth direction (*c* axis), confirming the screw character of the dislocation.

Further radial growth on those nanowires with the Eshelby twist results in mesoscale twisted GeS (Fig. 4a). We found that the gold nanoparticle that catalyses the VLS growth of the nanowire was present at the tip of the mesoscale structure (Fig. 4b), with a size proportional to the period of the structure; here, a large catalyst particle corresponds to a large period. The twist rates of the mesoscale structures are 0.2–1.1 rad μm^{−1}, comparable to the measured twist rate of the GeS nanowire. Further analysis established a good linear relationship between the overall twist rate of the mesoscale structure and the inverse

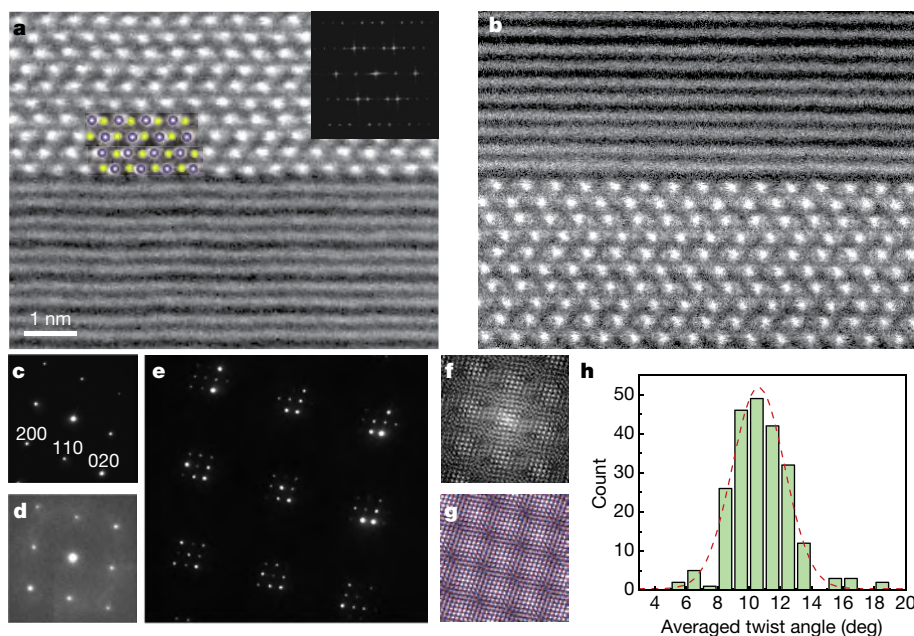


Fig. 2 | Twist interfaces and twist angles in mesoscale GeS structures. **a**, High-angle annular dark-field scanning transmission electron microscopy (HAADF-STEM) image showing the atomically sharp interface between two nanoplates in a twisted structure with the upper crystal on the [010] zone axis. A structural model of GeS with labelled Ge atoms (blue) and S atoms (yellow) is superimposed on the STEM image. The inset shows the fast Fourier transform (FFT) pattern of the upper crystal. **b**, HAADF-STEM image of the same interface with the lower crystal on the [010] zone axis. **c**, **d**, Selected area electron diffraction

(SAED) patterns for the [001] zone axis acquired from a plan-view TEM sample (see also Extended Data Fig. 3) containing two stacking nanoplates. SAED patterns were acquired on the top crystal (**c**) and bottom crystal (**d**), suggesting a misorientation angle of 7.5° between those two crystals. **e**, Double diffraction pattern acquired from the twist interface of the two nanoplates. **f**, **g**, Inverse FFT images of the double diffraction pattern (**f**) and the simulated rotational moiré pattern (**g**), illustrating the long-range periodicity in real space. **h**, Distribution of twist angle estimated by counting the number of interfaces per period using SEM imaging.

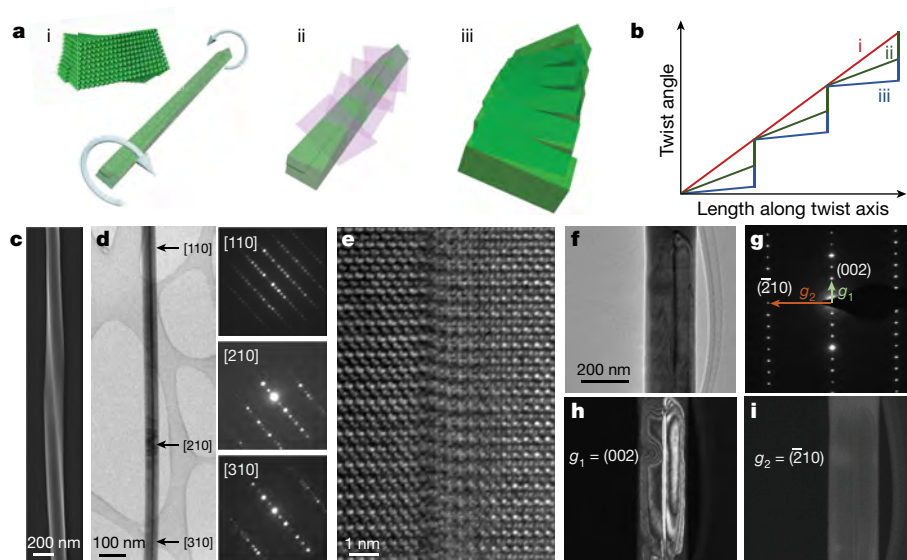


Fig. 3 | GeS nanowires with Eshelby twist. **a**, Schematics showing the formation mechanism of the mesoscale GeS structures with discrete twisting. **i**, The growth of nanowire with Eshelby twist. The inset shows a schematic of the atomic structure of the Eshelby twist. **ii**, Interlayer slip that forms the twist grain boundary. **iii**, Further radial growth giving rise to the discretely twisting morphology. **b**, Schematic diagram showing the twisting profile transitions from a continuous twist to a discrete twist, corresponding to states **i**–**iii** in **a**. **c**, SEM image of a twisted nanowire adhering to a substrate. **d**, TEM image of a nanowire with Eshelby twist (left) and corresponding CBED patterns (right) taken from different locations on the nanowires, suggesting that the crystal orientation changes

from direction $[110]$ to $[210]$ and then to $[310]$ with respect to the incident electron beam. The locations are marked with black arrows. Note the CBED patterns are rotated 38° anticlockwise with respect to the image of the nanowire. **e**, High-resolution TEM image of the dislocation. **f**, TEM image of another nanowire with axial dislocation tilted on the $[120]$ zone axis. **g**, Corresponding SAED pattern for the nanowire in **f**. The arrows in the SAED pattern highlight the reflections used for $\mathbf{g} \cdot \mathbf{b}$ analysis. **h**, **i**, Dark-field TEM images under the two-beam condition showing high (**h**) and low (**i**) contrast of the dislocation when $\mathbf{g}_1 = (002)$ and $\mathbf{g}_2 = (\bar{2}10)$ are excited, respectively.

of the cross-sectional area of the catalyst particle at its contact with GeS. The size of the catalyst particle at the tip of mesoscale structures approximately represents the radial size of the nanowires first axially grown by the VLS process (Fig. 4c). The relationship between the twist rate and catalyst size effectively reflects the relationship between the twist rate and cross-sectional area of the VLS-grown nanowires (Fig. 4f), revealing the Eshelby twist mechanism underpinning the formation of the discretely twisting structure. Fitting of the results using the Eshelby model gives a reasonable magnitude of 1.75 nm for κb , which is between one and two lattice constants along the c axis ($c = 1.04 \text{ nm}$). This magnitude agrees with the fact that the Burgers vector tends to be the shortest lattice vector (c) to reduce the elastic energy. For a more accurate evaluation of the Burgers vector from the Eshelby model, the anisotropic crystal structure and the complex cross-section of the nanowire should be considered, to correct the overestimation of area arising from our assumption of a circular cross-section and to determine the prefactor κ (ref. ¹⁵).

Although the Eshelby twist is essential, it cannot completely account for the discretization of the mesoscopic twist. For free-standing nanowires, radial growth increases the rigidity of the material, which counteracts the torque applied by the axial dislocation, resulting in the untwisting of the nanowire. This untwisting is in agreement with Eshelby's theory, which shows that the twist rate inversely scales with the cross-sectional area, so that at a radius of a few micrometres, in the Eshelby model, the twist rate becomes negligible¹⁵. Thus, the high twist rates in the mesoscale twisted GeS that are comparable to the twist rate of dislocated nanowires cannot be interpreted through the Eshelby model alone. The key to the formation of the discretized twisting in the structure is the interaction of the dislocated nanowire with the substrate. During growth, some of the free-standing dislocated nanowires adhere to the substrate. Unlike the free-standing nanowires, further radial growth of the nanowires pinned by the substrate does not result in untwisting, and the high twist rate of the initial nanowire would thus be preserved. This invariance in the twist rate with radial growth is the mechanism that permits us to detect the Eshelby twist in mesoscale GeS

(Fig. 4f). Our experiments (Extended Data Fig. 6) verify that the overall twist rate of a pinned structure remains almost constant with increasing radial size, and the twist rate of the mesoscale structure equals the twist rate of the initial nanowire upon substrate pinning.

Without the freedom to untwist, the volumetric elastic strain energy of the pinned nanowires rapidly builds up with radial growth. At a critical radial size, rotational slip of atomic layers is enabled by the weak interlayer vdW bonding in GeS to relieve the strain energy, resulting in the formation of twist boundaries and the discretization of the structure. The reduction of strain energy is counteracted by the increase in interfacial energy associated with the twisted interfaces, and this interplay defines the final twisting morphology. The elastic strain energy of the mesoscale structure is revealed by photoluminescence measurements (Fig. 4d, e). The photoluminescence emission peak is redshifted by 30 meV (from 748 nm to 766 nm) from the outer region, where the photoluminescence emission is consistent with the emission of strain-free GeS (1.65 eV)²⁵, to the central region of the structure. This result indicates the existence of torsional strain around the centre of the structure, in accordance with modification of bandgaps induced by screw dislocations as suggested in previous studies^{26,27}.

To illustrate this energy competition, we theoretically calculated the change in total energy, ΔE_{tot} , for each segment of the nanowire of length Δl , upon introducing a twist boundary into a pinned nanowire with a twist rate defined by the initial radius R_i of the nanowire upon its adhesion to the substrate (see Methods). We propose that ΔE_{tot} has two contributions, $\Delta E_{\text{tot}} = \Delta E_{\text{elastic}} + \Delta E_{\text{dis}}$, namely, a term $\Delta E_{\text{elastic}}$ due to the change in elastic energy in each segment of the wire with length Δl , and a term ΔE_{dis} due to the introduction of the misfit dislocations defining a twist boundary. The $\Delta E_{\text{elastic}}$ is found to be dependent on both the spacing between the twist boundaries Δl and the radial size of the pinned nanowire. Figure 4g displays ΔE_{tot} for a pinned nanowire with $R_i = 36 \text{ nm}$ as a function of Δl for three different values of radius. The result suggests a critical radius R_c , above which it becomes energetically favourable to introduce twist boundaries with a finite

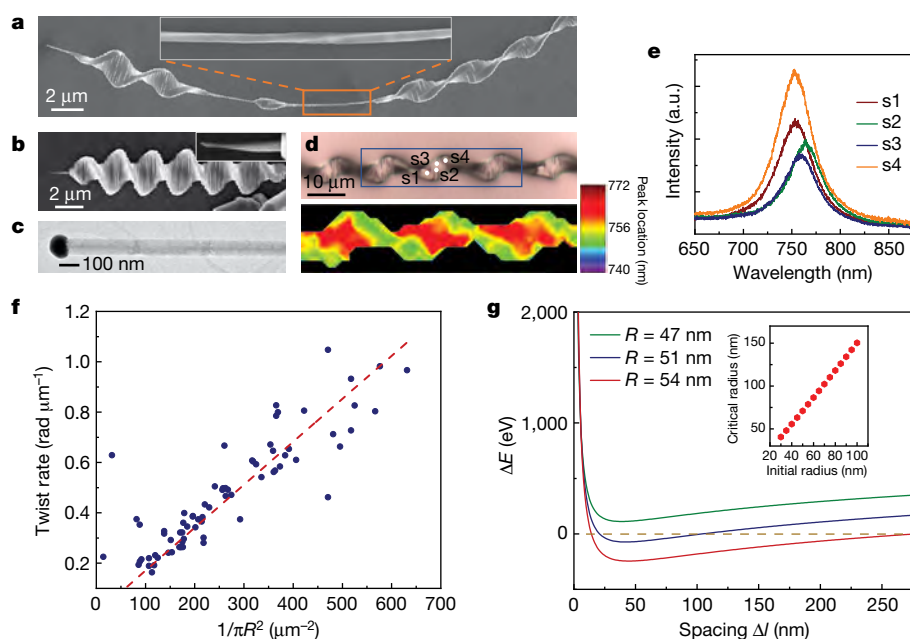


Fig. 4 | Formation mechanism of twisted GeS. **a**, SEM image of a structure with varying radial size that is formed by non-uniform growth of rotating nanoplates on a twisted nanowire. The inset of **a** highlights the twisting morphology of the nanowire. **b**, SEM image showing a gold nanoparticle at the tip of a mesoscale twisted structure. Inset, the high-magnification SEM image showing the gold nanoparticle (bright spot) at the tip. **c**, TEM image showing the gold catalyst (black particle) at the tip of a dislocated nanowire. **d**, Optical photo of a twisted GeS structure (top) and corresponding photoluminescence spectral maps of peak wavelength position (bottom). The blue frame shows the area for the

photoluminescence mapping. **e**, Photoluminescence spectra acquired on spots s1 to s4 marked in **d**, showing the variation of peak position. **f**, Scatterplot of twist rate measured on 73 individual twisted crystals as a function of the inversed contact areas $(\pi R^2)^{-1}$ between gold nanoparticles and the tip. The dashed red line is a least-squares linear fit through the data. **g**, The change in the total energy for a nanowire with twist rate defined at $R_i = 36$ nm, as a function of the spacing between the twist boundaries Δl , for three different values of nanowire radii. The inset shows the change of critical radius as a function of the initial radius of the nanowire, R_i , that defines the twist rate of the pinned structure.

spacing Δl . The critical radius depends on R_i , which defines the twist rate of the pinned nanowire (Fig. 4g); a small initial radius corresponding to a high twist rate results in a small critical radius. The radial size of dislocated nanowires for our growth ranges from 30 nm to 100 nm, giving rise to critical radii from 40 nm to 150 nm.

For further insight into the dependence of the topology on the radial size of the structure, we examined a twisted nanowire pinned on the substrate with a radius of around 150 nm. TEM analysis (Extended Data Fig. 7) confirms that the nanowire possesses twist boundaries at very small twist angles, as well as an almost continuous twisting profile. This very small discrete twisting at the boundaries combined with continuous twisting in the segments between the boundaries is an example of an intermediate twisting state at the onset of formation of the twist boundary. Thus, the 150 nm radial size of the nanowire approximates the critical radius to form the twist boundary. This is in reasonable agreement with the range of critical radii calculated from our theoretical model (Fig. 4g). The result also reveals the gradual transition of the twisting morphology from initial continuous twisting to intermediate twisting (consisting of both continuous twisting between the twist boundaries and discrete twisting at the boundaries) and eventually to discrete twisting with increasing radial size (schematically shown in Fig. 3a, b). This allows us to control the twisting profile and angles at twist interfaces by controlling the radial growth of the structure.

The growth mode identified in GeS is likely to be generic and could be used to produce twisted structure in other vdW materials. To demonstrate its versatility, we synthesized mesoscale germanium selenide (GeSe) structures with discrete twisting by depositing GeSe on the twisted GeS nanowires (Extended Data Fig. 8). In addition, the twisted structures can be transferred to other substrates using a simple wet-transfer method (Extended Data Fig. 9). The bottom-up scheme provides an approach to manipulate the twisting morphology and to realize defect engineering in vdW materials, providing new

opportunities to tailor their electrical, optical and thermal properties. Through controlling the radial growth, atomically sharp twist interfaces with a wide and tunable range of twist angles can be created, providing ways to explore ‘twistronic’ effects in these materials. In addition, the twisting period can be tailored by controlling the diameter of the dislocation-containing nanowires. In contrast to planar and twisted structures, axial structures formed by helical stacking of atomic layers or nanoplates can have a large and modulated chiroptical response due to the periodicity and increased interaction length with light. Moreover, the axial screw dislocation spirally threads the vdW layers, which can improve the electron conductivity along the cross-plane direction²³. The presence of the dislocation and the twist also enhance the scattering of phonons, lowering the thermal conductivity²⁸. The combination of increasing electron conductivity while decreasing thermal conductivity is attractive for thermoelectrics.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-019-1308-y>.

Received: 8 November 2018; Accepted: 26 April 2019;

Published online 19 June 2019.

- Kim, C.-J. et al. Chiral atomically thin films. *Nat. Nanotechnol.* **11**, 520–524 (2016).
- Song, J. C. & Gabor, N. M. Electron quantum metamaterials in van der Waals heterostructures. *Nat. Nanotechnol.* **13**, 986–993 (2018).
- Cao, Y. et al. Correlated insulator behaviour at half-filling in magic-angle graphene superlattices. *Nature* **556**, 80–84 (2018).
- Cao, Y. et al. Unconventional superconductivity in magic-angle graphene superlattices. *Nature* **556**, 43–50 (2018).
- Liu, K. et al. Evolution of interlayer coupling in twisted molybdenum disulfide bilayers. *Nat. Commun.* **5**, 4966 (2014).
- Naik, M. H. & Jain, M. Ultraflatbands and shear solitons in moiré patterns of twisted bilayer transition metal dichalcogenides. *Phys. Rev. Lett.* **121**, 266401 (2018).

7. Kang, P. et al. Moiré impurities in twisted bilayer black phosphorus: effects on the carrier mobility. *Phys. Rev. B* **96**, 195406 (2017).
8. Shtukenberg, A. G., Punin, Y. O., Gujral, A. & Kahr, B. Growth actuated bending and twisting of single crystals. *Angew. Chem. Int. Ed.* **53**, 672–699 (2014).
9. Jin, S., Bierman, M. J. & Morin, S. A. A new twist on nanowire formation: screw-dislocation-driven growth of nanowires and nanotubes. *J. Phys. Chem. Lett.* **1**, 1472–1480 (2010).
10. Bierman, M. J., Lau, Y. A., Kvit, A. V., Schmitt, A. L. & Jin, S. Dislocation-driven nanowire growth and Eshelby twist. *Science* **320**, 1060–1063 (2008).
11. Zhu, J. et al. Formation of chiral branched nanowires by the Eshelby twist. *Nat. Nanotechnol.* **3**, 477–481 (2008).
12. Oaki, Y. & Imai, H. Amplification of chirality from molecules into morphology of crystals through molecular recognition. *J. Am. Chem. Soc.* **126**, 9271–9275 (2004).
13. Feng, W. et al. Assembly of mesoscale helices with near-unity enantiomeric excess and light–matter interactions for chiral semiconductors. *Sci. Adv.* **3**, e1601159 (2017).
14. Srivastava, S. et al. Light-controlled self-assembly of semiconductor nanoparticles into twisted ribbons. *Science* **327**, 1355–1359 (2010).
15. Eshelby, J. The twist in a crystal whisker containing a dislocation. *Philos. Mag.* **3**, 440–447 (1958).
16. Eshelby, J. Screw dislocations in thin rods. *J. Appl. Phys.* **24**, 176–179 (1953).
17. Sutter, E. & Sutter, P. 1D wires of 2D layered materials: germanium sulfide nanowires as efficient light emitters. *ACS Appl. Nano Mater.* **1**, 1042–1049 (2018).
18. Li, C., Yu, Y., Chi, M. & Cao, L. Epitaxial nanosheet–nanowire heterostructures. *Nano Lett.* **13**, 948–953 (2013).
19. Kong, D. et al. Topological insulator nanowires and nanoribbons. *Nano Lett.* **10**, 329–333 (2010).
20. Peng, H., Xie, C., Schoen, D. T. & Cui, Y. Large anisotropy of electrical properties in layer-structured In_2Se_3 nanowires. *Nano Lett.* **8**, 1511–1516 (2008).
21. Zhang, L. et al. Three-dimensional spirals of atomic layered MoS_2 . *Nano Lett.* **14**, 6418–6423 (2014).
22. Shearer, M. J. et al. Complex and noncentrosymmetric stacking of layered metal dichalcogenide materials created by screw dislocations. *J. Am. Chem. Soc.* **139**, 3496–3504 (2017).
23. Ly, T. H. et al. Vertically conductive MoS_2 spiral pyramid. *Adv. Mater.* **28**, 7723–7728 (2016).
24. Tamura, N. in *Strain and Dislocation Gradients from Diffraction: Spatially Resolved Local Structure and Defects* (eds Barabash, R. & Ice, G.) 125–155 (Imperial College Press, 2014).
25. Tan, D. et al. Anisotropic optical and electronic properties of two-dimensional layered germanium sulfide. *Nano Res.* **10**, 546–555 (2017).
26. Akatyeva, E., Kou, L., Nikiforov, I., Frauenheim, T. & Dumitrica, T. Electrically active screw dislocations in helical ZnO and Si nanowires and nanotubes. *ACS Nano* **6**, 10042–10049 (2012).
27. Albrecht, M., Lympirakis, L. & Neugebauer, J. Origin of the unusually strong luminescence of a-type screw dislocations in GaN . *Phys. Rev. B* **90**, 241201 (2014).
28. Al-Ghalith, J., Ni, Y. & Dumitrică, T. Nanowires with dislocations for ultralow lattice thermal conductivity. *Phys. Chem. Chem. Phys.* **18**, 9888–9892 (2016).

Acknowledgements Y.L. and J.Y. are supported by the Samsung Advanced Institute of Technology under the grant 037361-003. Work at the Molecular Foundry and the Advanced Light Source was supported by the Office of Science, Office of Basic Energy Sciences, of the US Department of Energy under contract no. DE-AC02-05CH11231. H.S. and D.C.C. are supported by the US Department of Energy, Office of Science, Office of Basic Energy Sciences, Materials Sciences and Engineering Division under contract no. DE-AC02-244 05CH11231 within the Electronic Materials Program (KC1201). This work was performed, in part, at the Center for Nanoscale Materials, a US Department of Energy Office of Science User Facility, and supported by the US Department of Energy, Office of Science, under contract no. DE-AC02-06CH11357. We thank C. So, C. Song, X. Wang, S. Yan, K. Bustillo and C. V. Stan for help with the experiments.

Reviewer information Nature thanks Hua Zhang and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions Y.L. and J.Y. conceived the project and designed the experiments. S.K., Z.F., Y.L. and M.W. synthesized the samples. N.T. performed the X-ray analysis. J. Wang, F.Y., Y.L. and D.J. prepared samples for TEM study. Y.L., J. Wang, J. Wen, K.B.T., X.S. and M.C.S. worked on the TEM measurements. R.Z., Q.Y., J.T., R.O.R. and A.M.M. performed the EBSD analysis. H.S., B.Z.X. and D.C.C. carried out the theoretical calculations. Y.L. and J.Y. wrote the manuscript. All authors discussed the results and commented on the manuscript.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-019-1308-y>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-019-1308-y>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to J.Y.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

METHODS

Synthesis. Twisted GeS was synthesized using a horizontal tube furnace with a diameter of 1 inch. The temperature profile of the furnace was measured with a thermocouple. Extended Data Fig. 1 shows a schematic diagram of the furnace and the temperature profiles of the furnace for different heating temperatures at the source site. The substrates used for the synthesis were Si(100) substrates with natural oxidation and thermally oxidized Si(100) substrates with 300 nm silicon oxide. Gold catalyst, 3 nm thick, was deposited on those substrates and was patterned to form microbars in a width of tens of micrometres that are spaced by a few hundred micrometres, using photolithography and electron-beam evaporation deposition. Before the synthesis, the furnace was pumped down to a base pressure of 5 mTorr and flushed with argon gas blended with 4% hydrogen several times. In the synthesis, GeS powder (Sigma-Aldrich) was placed at the centre of the tube and heated to evaporate at a fixed pressure flowing with Ar/4% H₂ carrier gas. The substrates were placed 10–12 cm downstream from the GeS source. Typical growth conditions utilize pressures of 1–2 Torr, flow rates of 20–50 standard cubic centimetres per minute (sccm), source temperatures of 400–450 °C, deposition temperatures of 350–400 °C and a growth time of 20 min.

Characterization. *Laue X-ray microdiffraction analysis.* The crystallinity and orientation of the twisted structures were examined by scanning Laue X-ray microdiffraction (μ SXRD) with submicrometre spatial resolution on beamline 12.3.2 of the Advanced Light Source synchrotron at the Lawrence Berkeley National Laboratory. The sample was raster scanned with a 0.5 μ m step size and a Laue pattern collected at each step. The Laue patterns were then indexed using the XMAS software, providing orientation map of the sample.

TEM sample preparation. Cross-sectional TEM specimens were prepared from mesoscale and nanoscale twisted GeS structures growing on silicon substrate using a FEI Strata 235 dual beam focused ion beam (FIB) system and an FEI Nova 600 FIB. Low-energy argon ion milling at 900 eV was further used to minimize sidewall damage and to thin the specimen for electron transparency. Free-standing nanowires that grew vertically on the substrate were mechanically transferred to copper TEM grids for TEM analysis.

Electron microscopy. SEM imaging was performed using a Zeiss Gemini Ultra-55 Analytical SEM. Electron backscattering diffraction (EBSD) was performed in a FEI Strata 235 dual beam FIB at 20 keV using an equipped EBSD detector. EBSD orientation maps were generated and analysed using OIM software. TEM analysis was performed using a FEI TitanX, a JEOL 3010 in situ TEM, an Argonne Chromatic Aberration-corrected TEM (ACAT), and a FEI Titan 80-300 ST with both spherical and chromatic aberrations correctors. Atomic resolution HAADF-STEM images were acquired using the monochromated, aberration-corrected TEAM 0.5 TEM at the Lawrence Berkeley National Laboratory, operating at 200 keV.

Optical characterization. Photoluminescence measurements and mapping were conducted using a Horiba Jobin Yvon LabRAM ARAMIS automated scanning confocal Raman microscope under 532 nm excitation.

Theoretical model. The periodicity of the twist boundaries in the nanowires can be explained using a simple model rooted in the Eshelby twist¹⁶ and the associated strain energy, the fact that the nanowires are adhered to the substrate, and the notion of a critical thickness for misfit dislocations to arise.

We can build a semiquantitative model for the process to estimate the spacing of the twist boundaries. We assume that the nanowire can, at all times, be modelled as a cylindrical wire with radius R . The wire grows very rapidly in the initial stages, owing to the VLS process and the screw dislocation along the wire axis. The radius of the nanowire at this stage is assumed to be R_i . Growth to radii beyond that of the gold droplet, however, is much slower, and is mediated by direct deposition. The screw dislocation along the axis of the nanowire introduces a net torsion into the wire, and this torsion leads to the so-called Eshelby twist¹⁶, where the twist rate of the wire, α , is given by:

$$\alpha = \frac{b_a}{\pi R_i^2} \quad (1)$$

where b_a is the Burgers vector of the axial screw dislocation.

In the initial growth process, the wire comes into contact with the substrate and bonds to it. The result of this bonding is that the total twist rate of the wire is fixed at α throughout the remainder of the growth process. As the radius of the wire increases during the growth, the equilibrium twist rate is reduced below that of equation (1). However, since the wire has a net twist rate given by equation (1), this leads to an excess of torsional strain energy stored in the wire. We hypothesize that this excess strain energy can be reduced by the introduction of twist boundaries into the nanowire. These twist boundaries can reduce the torsional strain energy in the cylinder but introduce an array of misfit dislocations to define the twist boundary. The interplay between these two effects can define the critical thickness separating twist boundaries²⁹. The critical distance so defined represents a lower

limit on the spacing of the twist boundaries, as kinetic effects might lead to larger spacings than those predicted by the model.

We will first assume that we can model the nanowire as a cylinder of radius R that is growing along the z -direction. The total change in elastic energy upon introducing twist boundaries with a spacing Δl can be computed. The twist boundaries are defined by two perpendicular arrays of screw dislocations that are introduced into the wire in crossed pairs. The introduction of n pairs of dislocations, each with Burgers vector of length b_m , at a spacing $d = \frac{2R}{n}$ within the plane of the twist boundary creates a twist boundary with the twist angle, θ , given by³⁰:

$$\theta = \frac{b_m}{d} = \frac{b_m n}{2R} \quad (2)$$

We then note that the change in elastic energy $\Delta E_{\text{elastic}}$ in a wire of length Δl , upon introducing the twist boundary described by equation (2) is given by:

$$\Delta E_{\text{elastic}} = -n \frac{b_a b_m K_s R^3}{4R_i^2} + n^2 \frac{b_m^2 K_s \pi R^2}{16 \Delta l} \quad (3)$$

Here K_s is the torsional shear modulus of the material.

The reduction in elastic energy, equation (3), is countered by the introduction of n pairs of misfit dislocations. The energy of these dislocations E_{dis} is taken to be, approximately,

$$E_{\text{dis}} = n \frac{R}{\pi} K_c b_m^2 \log \left(\frac{\Delta l}{\beta b_m} \right) \quad (4)$$

with βb_m the core radius of the misfit dislocations defining the twist boundary, and K_c the elastic constant governing the line energy of the dislocation computed from anisotropic elasticity theory^{30–32}. For simplicity, we take both sets of screw dislocations defining the twist boundary to have the same Burgers vector, and we assume that all have the same length, $2R$. (GeS is orthorhombic, so the screw dislocations in question have slightly different lengths of Burgers vectors. We choose the shorter of the two vectors, and the spacing is then that associated with the smallest of the Burgers vectors. For a pure twist boundary, the ratio b_m/d is fixed to be the same constant for both sets of dislocations.)

These two contributions can be summed to give the total change in energy upon introduction of a twist boundary with n pairs of misfit dislocations²⁹ at intervals of Δl . We choose to measure all lengths in terms of b_m , all elastic constants in terms of K_c and all energies in units of $K_s b_m^3$. In non-dimensional form, the expression for the change of energy, ΔE_{tot} , for a wire of radius R with twist rate fixed at a radius R_i , becomes:

$$\Delta E_{\text{tot}} = \frac{1}{16} n R \left[R \left(-\frac{4b_a R}{R_i^2} + \frac{n\pi}{\Delta l} \right) + \frac{16}{\pi} K_c \log \left[\frac{\Delta l}{\beta} \right] \right] \quad (5)$$

where all variables now refer to their dimensionless versions.

Equation (5) enables exploration of the relationship between twist boundary spacing and the materials properties. There are only three constants that appear in equation (5): β , b_a and K_c . The parameter β defines the core radius of the misfit dislocations. Typically, one expects β to be the order of 1, and we choose to set $\beta = 1$. The in-plane lattice parameters of GeS (according to the Materials Project Database³³) are given by $a = 0.36$ nm and $b = 0.44$ nm. We choose a to set our length scale. With this choice, the dimensionless axial Burgers vector length becomes $b_a = 2.889$.

We begin by considering ΔE_{tot} as a function of Δl , at fixed values of R_i and R . Determination of the exact value for K_c is beyond the scope of this project. On the basis of rough calculations, we estimate that $K_c = 3/4$. Extended Data Fig. 10 displays the change in total energy for a cylinder with dimensionless twist radius defined at radius $R_i = 100$ (that is, 36 nm, comparable to the experiment) for three different values of R . Note that there is a critical value for R at which it becomes possible to introduce the misfit dislocations. When R exceeds this value, the dislocations can be introduced over a finite range of Δl . This unusual dependence is simple to understand. For small Δl , the twist rate is markedly decreased through the introduction of a single twist boundary. However, as Δl grows, the average twist rate reduction due to each twist boundary decreases. The result is that if Δl is too large, the reduction in torsional strain energy due to a single twist boundary cannot be large enough to generate the dislocations of the twist boundary.

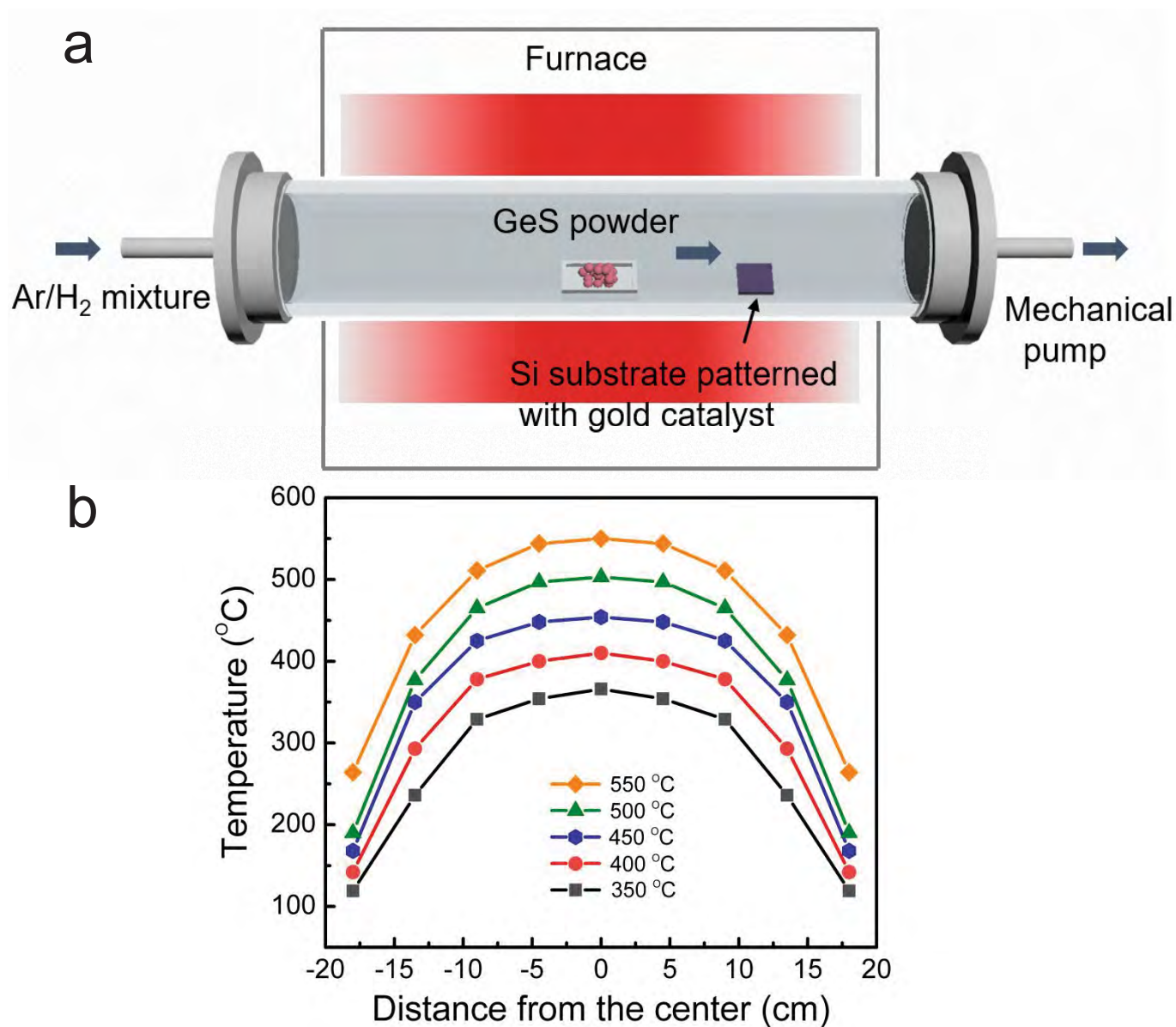
The critical Δl values predicted by the model at the critical R are too small, hovering near 30 nm, whereas the experiment indicates values nearer to 200 nm. There are many potential origins for this discrepancy, but we believe that the kinetics of misfit dislocation introduction are the most important. Consider a cylinder, again with the twist radius initially fixed at $R_i = 100$ (in unit of b_m). As it grows below the critical radius, there is no driving force to create any twist boundary, since introducing a twist boundary always increases the energy. However, as the cylinder slightly exceeds the critical radius (for example $R = 140$), a twist bound-

ary can be introduced, and for a larger radius ($R = 150$), it is possible to have separations up to approximately 800 (in dimensionless units) between twist boundaries while still reducing the total energy. On reinserting dimensions into these equations, the model predicts that twist boundaries can be introduced into the nanowire when its radius approaches 50 nm. By the time the wire is 54 nm in radius, the critical spacing range over which a twist boundary would be stable, Δl , includes thicknesses ranging from 18 nm to 288 nm. It is likely that introduction of the misfit dislocations to create the twist boundaries will require the overcoming of kinetic limitations, and that the nanowire will continue to grow radially while these kinetic limitations are being overcome. Given the sensitivity to the radius of the wire and the simplicity of the model, the separations agree well with the experimental observations.

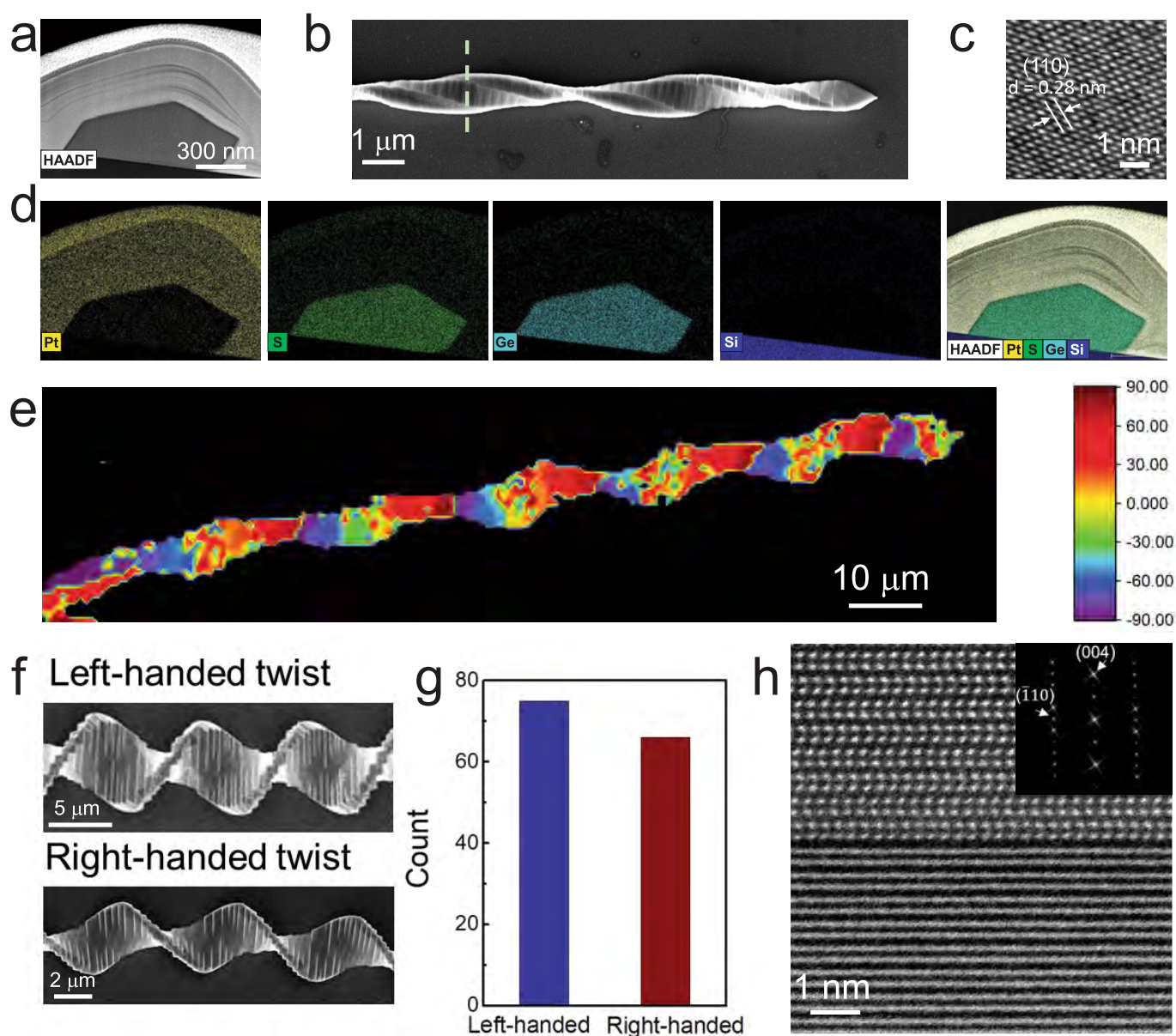
Data availability

All data supporting the findings of this study are available within the paper.

29. Ertekin E., Greaney, P. A., Chrzan, D. C. & Sands, T. D. Equilibrium limits of coherency in strained nanowire heterostructures. *J. Appl. Phys.* **97**, 114325 (2005).
30. Hirth, J. P. & Lothe, J. *Theory of Dislocations* (Krieger, 1992).
31. Eshelby, J. D., Read, W. T. & Schockley, W. Anisotropic elasticity with applications to dislocation theory. *Acta Metall.* **1**, 251–259 (1953).
32. Foreman, A. J. E. Dislocation energies in anisotropic crystals. *Acta Metall.* **3**, 322–330 (1955).
33. de Jong, M. et al. Charting the complete elastic properties of inorganic crystalline compounds. *Sci. Data* **2**, 150009 (2015).

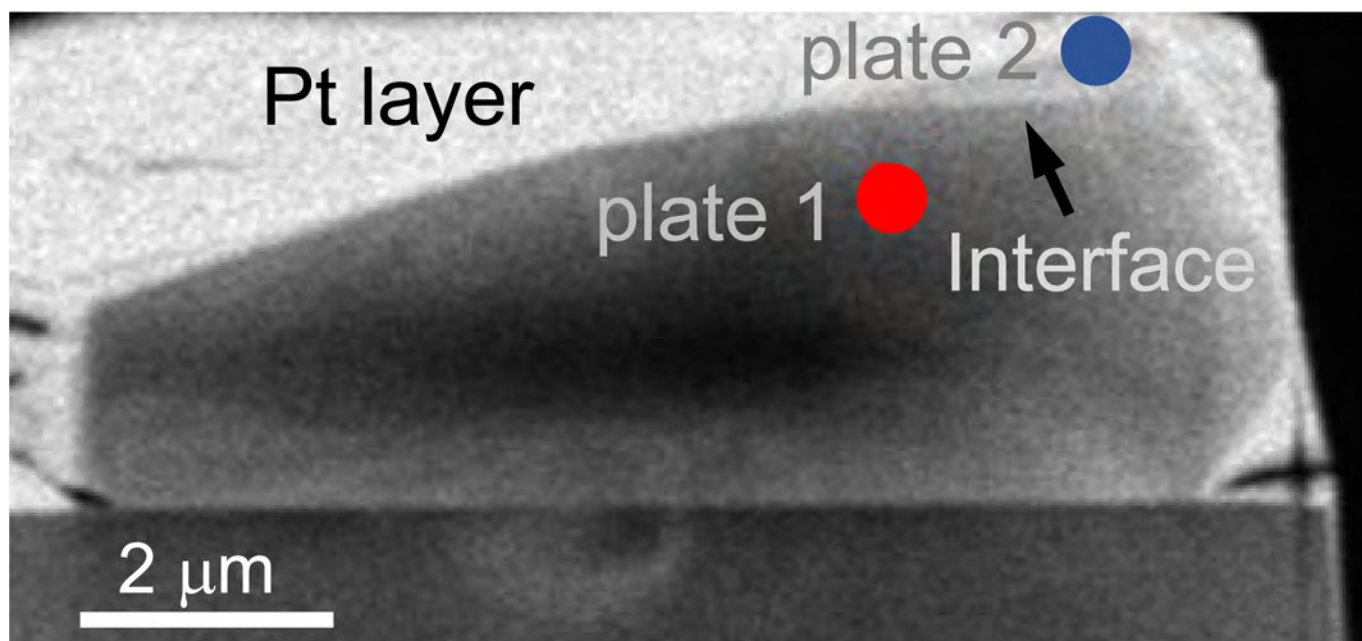


Extended Data Fig. 1 | Furnace set-up for the growth. **a**, Schematic diagram of the furnace used for the synthesis of the twisted GeS crystals. **b**, Temperature profiles of the furnace for heating temperatures of 350 °C, 400 °C, 450 °C, 500 °C and 550 °C at the source site.



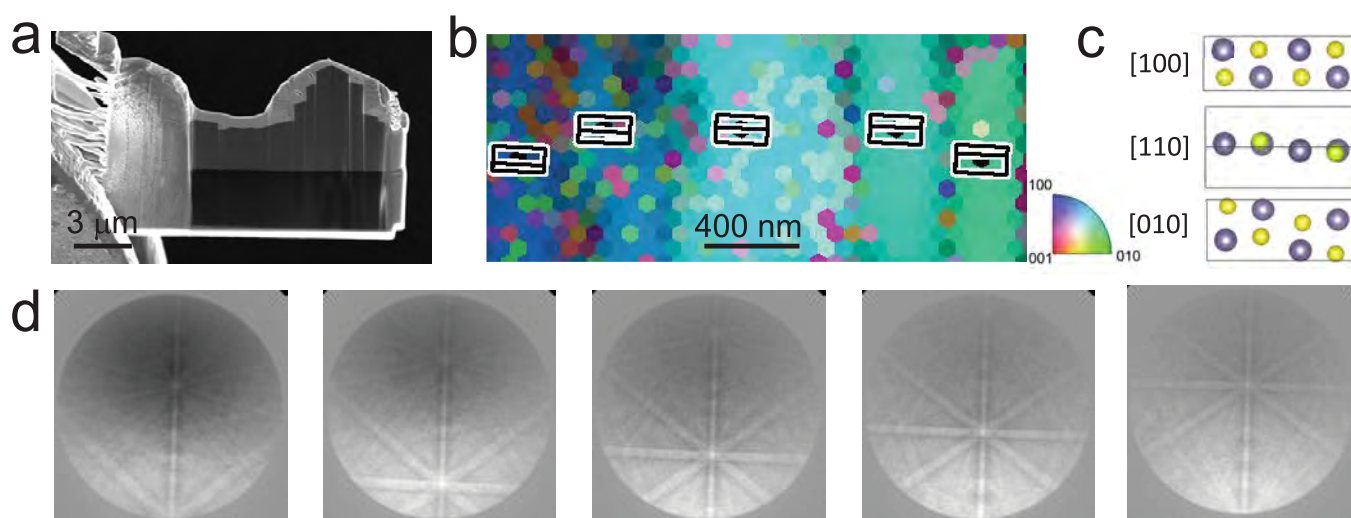
Extended Data Fig. 2 | Additional chemical and structural analysis of mesoscale twisted GeS. **a–d**, Chemical analysis of mesoscale twisted GeS using energy-dispersive X-ray spectroscopy (EDS). **a**, HAADF-STEM image of a cross-sectional lamellar sample with its normal perpendicular to the twist axis prepared by FIB milling. **b**, SEM image showing the crystal used to prepare the TEM sample. The dashed line represents the location of the cross-section. **c**, High-resolution TEM (HRTEM) image of the cross-section confirming that the twist axis is aligned with the [001] direction. **d**, STEM-EDS elemental map of the cross-section verifying that the structure is a compound consisting of Ge and S in an atomic ratio of 1:1. **e**, Additional X-ray microdiffraction analysis on a twisted GeS crystal with a period of about 15 μm . The orientation angle in the X-ray orientation map is defined by the angle between the b axis and the normal to the substrate. The large period, corresponding to a low crystallographic twist,

facilitates the use of X-ray microdiffraction to determine the crystal orientation. The X-ray analysis also suggests that the widest portion in a period has its b axis aligned with the substrate normal while the narrowest portion in the period has its a axis aligned with the substrate normal. This growth phenomenon may result from the structural anisotropy of GeS. **f, g**, Handedness of mesoscale twisted GeS. **f**, Representative SEM micrographs showing twisted structures with opposite helicity. **g**, Histogram showing that the population of left-handed structures is approximately equal to the population of right-handed structures. The measured ratio of left-handedness to right-handedness is 74:67. **h**, Additional atomic-resolution HAADF-STEM image of a twist interface in a twisted GeS structure with the upper crystal oriented on the [110] zone axis. The inset shows the FFT pattern of the upper crystal.



Extended Data Fig. 3 | TEM image of the specimen on which diffraction patterns in Fig. 2c–e were obtained. Using FIB, a lamellae sample containing two adjacent stacking nanoplates was prepared with the surface normal along the twist axis. After the thinning of the sample, the largest fraction of the sample consists of a single nanoplate (either the top or the

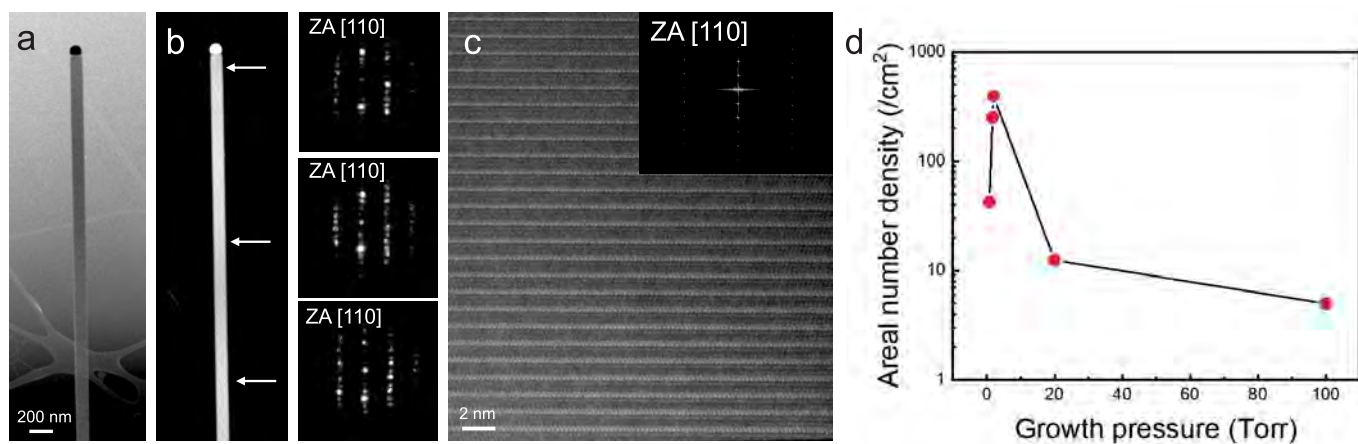
bottom one), while a smaller fraction preserves the twist interface between the two nanoplates. The arrow shows the location of the twist interface. The red and blue dots show the locations that respectively contain the single upper plate and the single lower plate.



Extended Data Fig. 4 | Electron backscattering diffraction analysis.

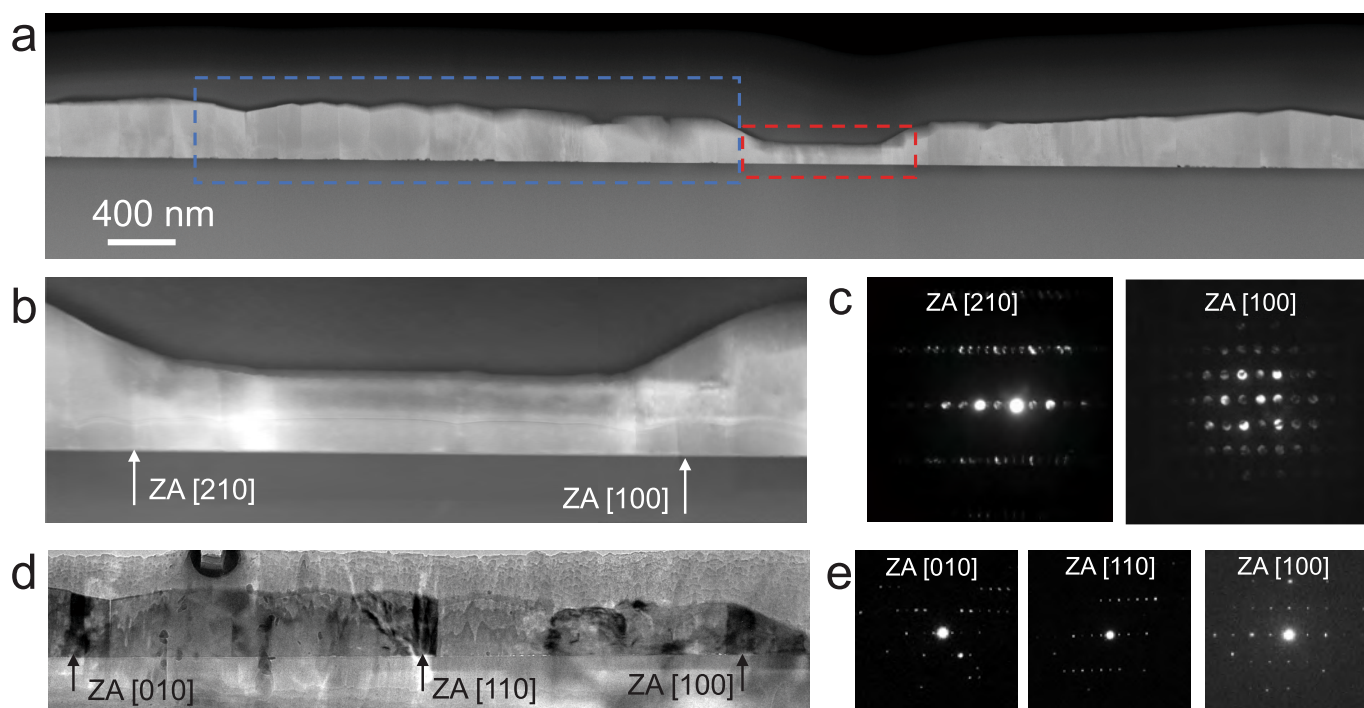
a, SEM image of a specimen used for EBSD. The lamellar sample is prepared by FIB milling with surface normal perpendicular to the twist axis of the structure. **b**, Representative EBSD orientation map of the nanoplates with the 2D projection of the unit cell superimposed. The unit cell projection shows the crystal orientation of the nanoplates at that point. The misorientations of the plates (that is, the differences between the crystal orientations of adjacent nanoplates) are quantified to be 10.2°, 14°, 6.7° and 8.7°. **c**, Unit cell of GeS viewed along the [100], [110] and [010]

directions. **d**, Corresponding EBSD patterns acquired from five adjacent nanoplates. Note that 13 twist angles were measured on four mesoscale twisted structures using EBSD, and a representative measurement is shown in this figure. In addition, two twist angles were measured using the TEM (shown in Fig. 2). In total, 15 twist angles were measured on six twist structures. The values of the measured twist angles are 10.6°, 16°, 6.8°, 10.3°, 13.9°, 7.1°, 8.3°, 14°, 9.6°, 10.2°, 14°, 6.7°, 8.7°, 10.27° and 7.5°. This is a range from 6.8° to 16° with an average of 10.3° and standard deviation of 3°.



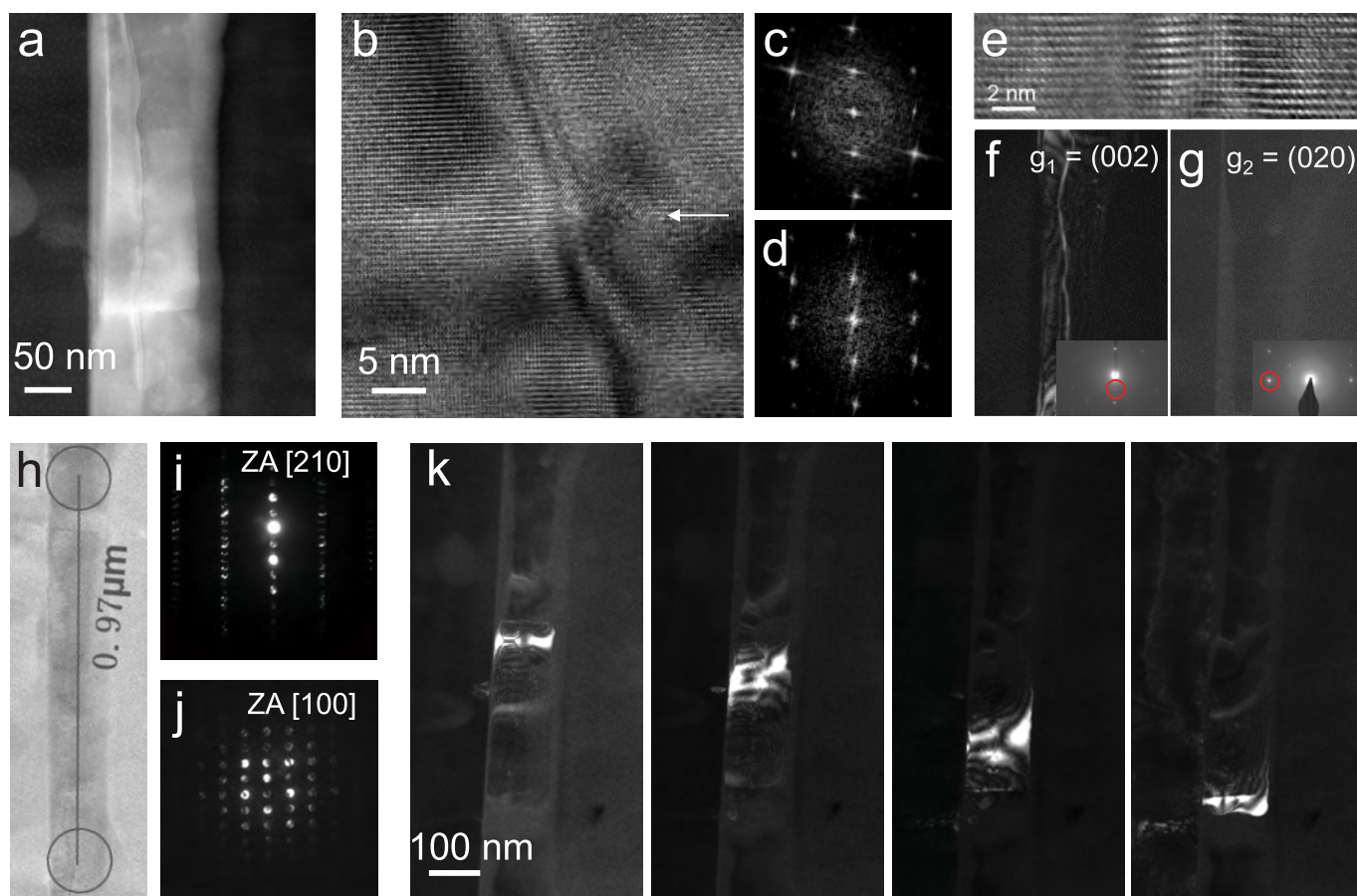
Extended Data Fig. 5 | The effect of growth pressure on the growth of dislocated nanowires. **a**, TEM image of a normal nanowire without a dislocation produced at a growth pressure of 5 Torr. **b**, Corresponding STEM image of the nanowire (left) and CBED patterns acquired from three different locations on the nanowire (right), showing the absence of the Eshelby twist. The white arrows in the STEM images show the locations where the CBED patterns are collected. **c**, HRTEM image of the normal nanowire showing that the growth direction of the nanowire is along the c axis. The inset is the FFT pattern of the HRTEM image

suggesting the image is taken on the [110] zone axis (ZA). **d**, Areal number density (the number of twisted structures per unit area of the substrate) of twisted GeS structures as a function of the growth pressure. In this experiment, the flow rate of Ar/H₂ carrier gas and the source temperature were fixed at 50 sccm and 450 °C, respectively. The growth of dislocated nanowires is achieved with typical growth pressures in the range 1–2 Torr and a flow rate of 20–50 sccm, whereas the yield of nanoscale and mesoscale twisted GeS drops when the growth pressure deviates from optimum.



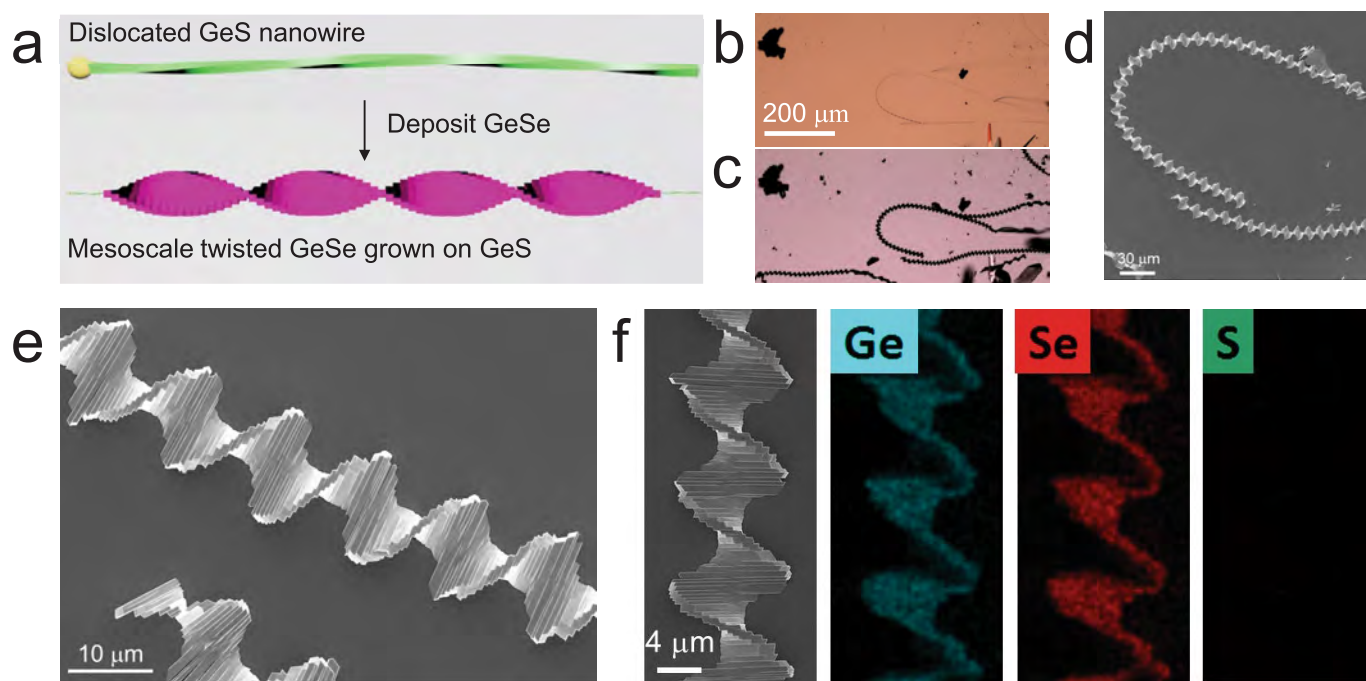
Extended Data Fig. 6 | The invariance of the total twist rate in the radial growth of a pinned structure. **a**, STEM image of a cross-sectional sample prepared from a twisted structure pinned on a substrate. This twist structure was formed by non-uniform radial growth on a twisted nanowire (similar to the structure shown in Fig. 4a), showing varying radial sizes for different portions of the structure. To verify the invariance of the total twist rate in the radial growth of a pinned structure, we used electron diffraction to measure the twist rates of different portions with varying radial sizes in the structure. **b**, Magnified STEM image of a thin portion highlighted by the dashed red box in **a**. **c**, CBED patterns for the [210] and [100] zone axes collected at locations marked by arrows in **b**. This suggests that the thin portion has a twist of 23° over a length of 970 nm, amounting to a twist rate of 0.4 rad mm^{-1} . **d**, Magnified TEM image showing a thick

portion highlighted by the dashed blue box in **a**. **e**, SAED patterns for the [010], [110] and [100] zone axes collected at locations marked by arrows in **d**. This suggests that the thick portion has a twist of 90° over a length of $4 \mu\text{m}$, giving rise to a twist rate of 0.4 rad mm^{-1} as well. This electron diffraction analysis shows that twist rates at different portions with varying radial sizes in the structure are almost same, despite the significant difference in their radial size. Note that the thick portion has a radial size (about 450 nm) two times larger than that of the thin portion (about 150 nm). This result suggests that the overall twist rate of the structure is determined by the twist rate of the nanowire upon substrate pinning, and further radial growth does not result in untwisting that decreases the twist rate. The high twist rate of the initial nanowire with Eshelby twist is therefore preserved during radial growth.



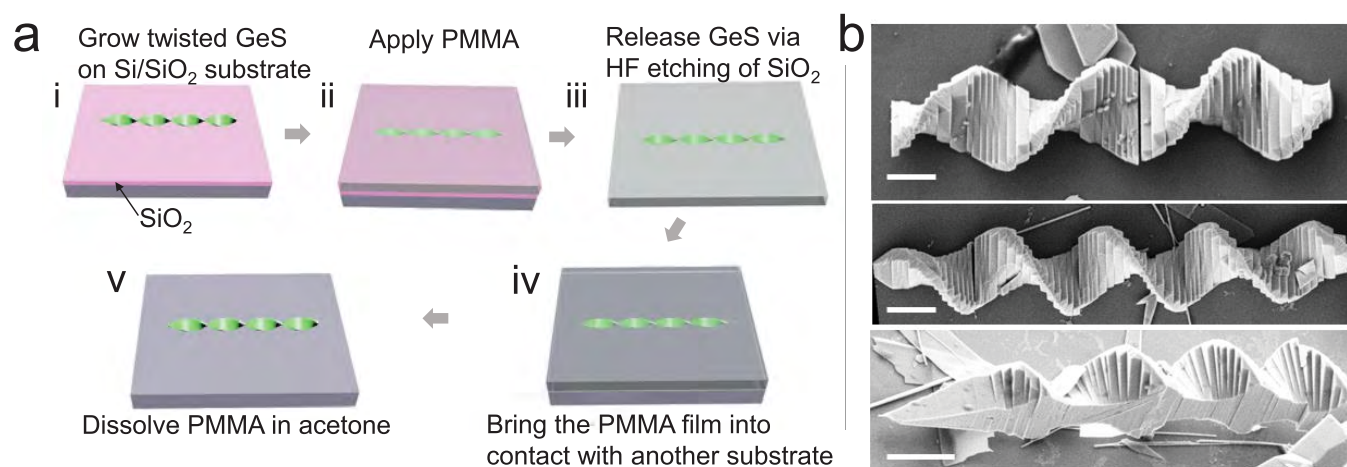
Extended Data Fig. 7 | Twisted GeS nanostructure in an intermediate twisting state. **a**, Low-magnification STEM image of a nanostructure growing horizontally on a substrate. The nanowire had an approximate radial size of 200 nm and height of 150 nm with a twisting morphology that can be clearly observed by SEM imaging. We note that this nanostructure is the thin part of the sample shown in Extended Data Fig. 6. In contrast to free-standing nanowires (Fig. 3d) that have only a screw dislocation, the nanowire was segmented with the presence of both transverse boundaries and a dislocation line in the middle. **b**, HRTEM view of a boundary. The white arrow shows the boundary. **c**, **d**, FFT pattern of the two crystals across the boundary. The HRTEM imaging and the corresponding FFT patterns confirm that the crystals across the boundary have almost the same orientation and thus the boundary takes on a very small twist angle. **e**, HRTEM image of the screw dislocation. **f**, **g**, Burgers vector analysis, based on the $g \cdot b$ contrast. To perform the analysis, the nanowire was first tilted to the [100] zone axis. Next, the sample was further tilted to create two-beam conditions for different diffraction spots in the diffraction pattern. Dark-field images of the dislocation were taken for $g = (002)$ (**f**) and $g = (020)$ (**g**). The insets

show the excitation of reflections for the dark-field imaging in which the selected reflections are marked with red circles. For $g = (002)$, high contrast of the dislocation is observed in the dark-field image (**f**), whereas for $g = (020)$, the dislocation becomes invisible in the image (**g**). We have therefore determined the Burgers vector of the dislocation to be along the [001] direction, which is the same as for the dislocated nanowires that were grown vertical and free-standing. **h**, Low-magnification cross-sectional TEM image of the nanowire. **i**, **j**, CBED patterns for the [210] and [100] zone axes are collected at locations marked by circles in **a**, which were separated by 970 nm. This suggests a twist of 23° about the c axis within this length, amounting to a twist rate of 0.4 rad mm^{-1} , which is comparable to the twist rate of mesoscale crystals. **k**, A series of dark-field TEM images showing that the [020] diffraction band progressively shifts when the sample is continuously rotated about its twist axis by tilting the TEM holder; this dark-field imaging verifies that the crystallographic twist of the nanowire is almost continuous. As such, the nanostructure has both twist boundaries with very small twist angles and an almost continuous twisting profile, exemplifying an intermediate twisting state at the onset of formation of the twist boundary.



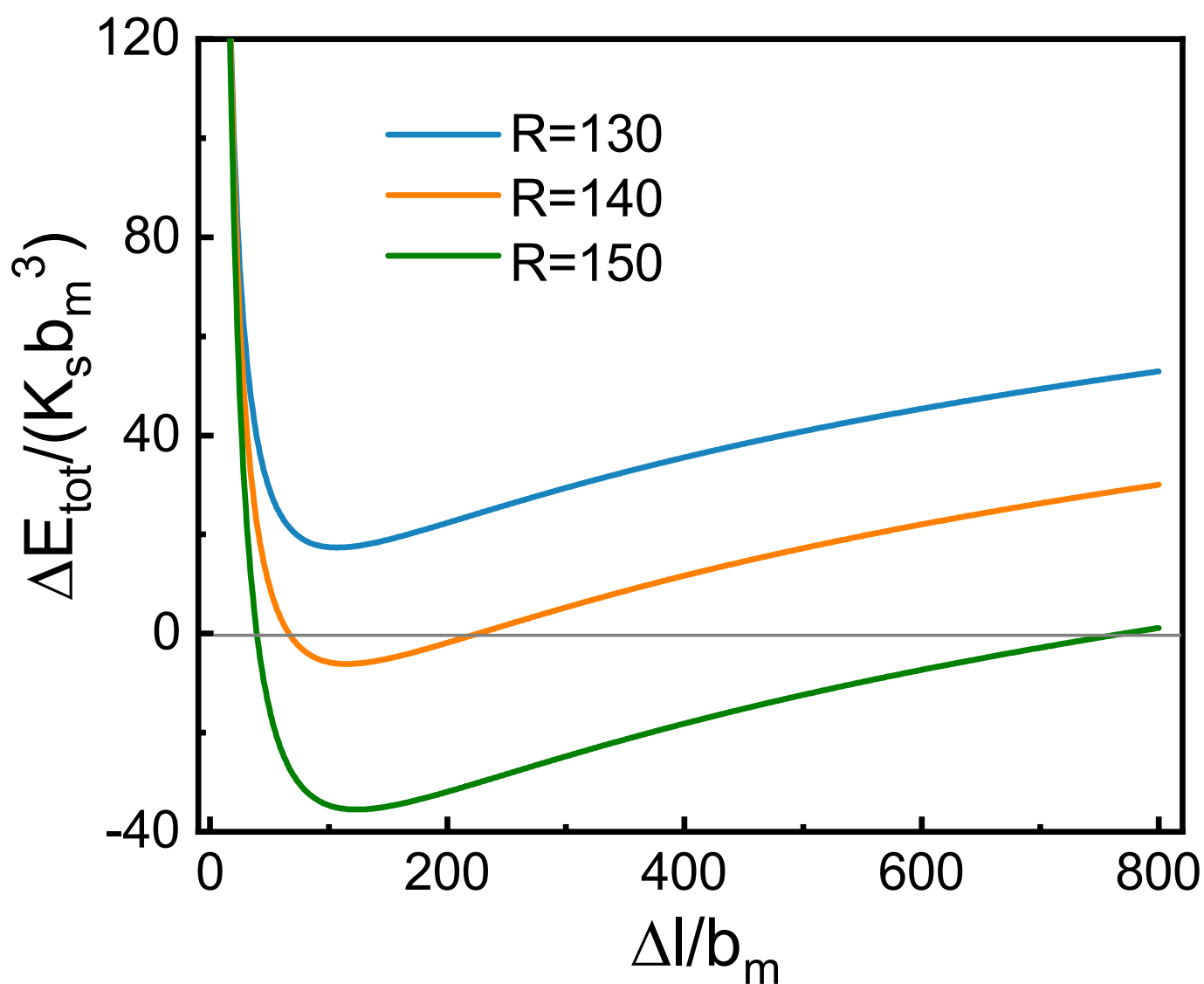
Extended Data Fig. 8 | Synthesis of mesoscale twisted GeSe using dislocated GeS nanowires as seeds. **a**, Schematic showing the synthesis of mesoscale twisted GeSe structures. Twisted GeS nanowires were first grown via the VLS method. In a second growth, GeSe was deposited on the GeS nanowires using the chemical vapour transport method. **b**, Optical image of dislocated GeS nanowires. **c**, Optical image of

mesoscale twisted structures synthesized through depositing GeSe on those twisted GeS nanowires. **d**, **e**, SEM images of twisted GeSe structures at low magnification (**d**) and at high magnification (**e**). **f**, SEM image (left) and corresponding EDS elemental maps of the structure. Quantitative chemical analysis using EDS suggests an almost 1:1 atomic ratio of Ge to Se.



Extended Data Fig. 9 | Transfer of twisted GeS crystals to other substrates. **a**, Schematic showing a facile processing scheme to transfer twisted GeS crystals to other substrates. i, Twisted crystals were first grown on a thermally oxidized Si/SiO₂ substrate. These crystals adhered well to the substrate. ii, Polymethyl methacrylate (PMMA) was applied on the

substrate. iii, The SiO₂ layer was etched using hydrofluoric acid, and the crystals were transferred to the PMMA film. iv, The PMMA film with the crystals was brought into contact with another substrate. v, The GeS crystals were transferred to the substrate by dissolving PMMA in acetone. **b**, SEM images of the GeS crystals after the transfer. Scale bar, 4 μ m.



Extended Data Fig. 10 | Change in energy. The total change in energy upon introduction of one dislocation pair into a nanowire of radius R (dimensionless), given that the initial twist rate of the wire is set at $R_i = 100$. Note that for these conditions there is a critical value of

R necessary to introduce misfit dislocations, as well as a critical thickness. Note also that the energy is only reduced over a range of Δl . See Methods for details.

Structural colour using organized microfibrillation in glassy polymer films

Masateru M. Ito^{1,2,4*}, Andrew H. Gibbons^{1,3,4}, Detao Qin^{1,2}, Daisuke Yamamoto¹, Handong Jiang^{1,2}, Daisuke Yamaguchi^{1,2}, Koichiro Tanaka^{1,3} & Easan Sivaniah^{1,2*}

The formation of microscopic cavities and microfibrils at stress hotspots in polymers is typically undesirable and is a contributor to material failure. This type of stress crazing is accelerated by solvents that are typically weak enough not to dissolve the polymer substantially, but which permeate and plasticize the polymer to facilitate the cavity and microfibril formation process^{1–3}. Here we show that microfibril and cavity formation in polymer films can be controlled and harnessed using standing-wave optics to design a periodic stress field within the film⁴. We can then develop the periodic stress field with a weak solvent to create alternating layers of cavity and microfibril-filled polymers, in a process that we call organized stress microfibrillation. These multi-layered porous structures show structural colour across the full visible spectrum, and the colour can be tuned by varying the temperature and solvent conditions under which the films are developed. By further use of standard lithographic and masking tools, the organized stress microfibrillation process becomes an inkless, large-scale colour printing process generating images at resolutions of up to 14,000 dots per inch on a number of flexible and transparent formats^{5,6}.

If a standing-wave light pattern is formed in a photosensitive polymer film, then the polymer will be selectively crosslinked in alternating layers in the film (Extended Data Fig. 1). Crosslinking of these layers leads to residual stresses that build across the non-crosslinked layers of the polymer. From our observations, we deduce that when such a film is exposed to a weak solvent, the stresses acting on the non-crosslinked layers will trigger microfibril formation to generate porous layers in the non-crosslinked regions (Fig. 1a), and an overall alternation of these microfibril layers with compact polymer layers.

This structure can be obtained in polystyrene thin films spincoated on silicon wafers that have been crosslinked under high-energy ultraviolet light. Spin-casting is known to leave residual stress in the thin films; these polystyrene films are de-stressed by annealing at high temperature (190 °C) before any ultraviolet exposure. When the exposed film is submerged in acetic acid, which acts as a weak solvent for polystyrene, stepwise changes in film colour are observed as the layers of microfibrils form (Extended Data Fig. 2).

Understanding this structure formation process during immersion in acetic acid is key to using it in future applications. Through scanning electron microscope (SEM) snapshots of the structure, the upper parts of the film are first seen to expand, until a fully open multilayer structure is produced. However, with further exposure to solvent, the upper layers begin to collapse until the entire film is fully closed again (Fig. 1b–e).

The most open form of the structure is highly periodic and composed of alternating dense and porous layers (Fig. 1c). The multi-layered microfibrils provide the useful property of structural colour, as observed by a Bragg peak in the film spectrum (Extended Data Fig. 3a)⁷. Optical transfer matrix analysis verifies the thicknesses of the photonic multilayer structure, 52 nm and 55 nm for the solid and microfibril layers respectively, and estimates a microfibril layer porosity

of 55% (ref. ⁸). Given that the standing-wave periodicity within the original film, 72 nm, is less than the final structure periodicity, this shows that expansion of the film has occurred during the acetic acid treatment.

This process can be followed by real-time spectroscopy measurements, indicating a characteristic structure formation behaviour during immersion in acetic acid (see Extended Data Fig. 3b–e). The real-time spectra can be analysed to determine the expansion of film thickness under three characteristic conditions (Fig. 1f, Extended Data Fig. 3b–d). During exposure to acetic acid at low temperatures (20 °C), the solvent permeates through each stressed but un-crosslinked layer, plasticizing it, and triggering abrupt microfibrillation of the layer; resulting in sudden increases in the overall film thickness. A staircase shape is observed in the expansion ratio with a step for each layer formed as the acetic acid diffuses from the top of the film to the bottom. If the film is developed at a higher temperature (30 °C), there is an additional relaxation of the plasticized microfibril structures. This leads to the subsequent collapse of the microfibril layers and a decrease in the film thickness. On the other hand, if the crosslinked polymer film is then thermally de-stressed by annealing at 160 °C (well above the glass transition temperature of polystyrene), there is no characteristic stepwise expansion in the film, and it is seen to slowly swell to levels seen in thin films with collapsed microfibrils⁹.

This last observation allows us to identify whether the organized stress microfibrillation process arises from differential stress or from differential swelling caused by the absorption differences between the crosslinked and less-crosslinked layers¹⁰. There is a large difference in the maximum expansion and the expansion rate of the film, depending on whether the thermal de-stressing is carried out below or above the glass transition, indicating that the crosslinking stress has been removed (Fig. 1g, h, Extended Data Fig. 3e, f). Moreover, the characteristic stepwise changes in film expansion disappear upon thermal removal of the film stress. The role of differential swelling cannot be completely discounted, because the acetic acid encounters differently crosslinked layers of polystyrene. However, if we were to assume that the microfibril structures were predominantly caused by differential swelling, it is difficult to explain the observations. In its broadest terms, the fundamental process behind organized stress microfibrillation is the stepwise, solvent-mediated expansion of a sequence of pre-stressed glassy polymer layers. By adjustment of the solvent conditions or the development time, it is possible to prevent subsequent collapse of the layers.

The layer periodicity is a standing-wave phenomenon, so we can alter it by crosslinking with different wavelengths of light. However, polystyrene is photosensitive only below 280 nm (ref. ¹¹). To expand polystyrene's photocrosslinking range, photoinitiators such as phenanthrenequinone (PQ) or 4,4'-bis(diethylamino) benzophenone (BDABP) were mixed into the polymer solution before thin film casting. These modified films were then crosslinked under longer-wavelength light-emitting diode (LED) light, up to 405 nm. We verified the role

¹Institute for Integrated Cell-Material Sciences (iCeMS), Kyoto University, Kyoto, Japan. ²Department of Molecular Engineering, Kyoto University, Kyoto, Japan. ³Department of Physics, Kyoto University, Kyoto, Japan. ⁴These authors contributed equally: Masateru M. Ito, Andrew H. Gibbons. *e-mail: mito@icems.kyoto-u.ac.jp; esivaniah@icems.kyoto-u.ac.jp

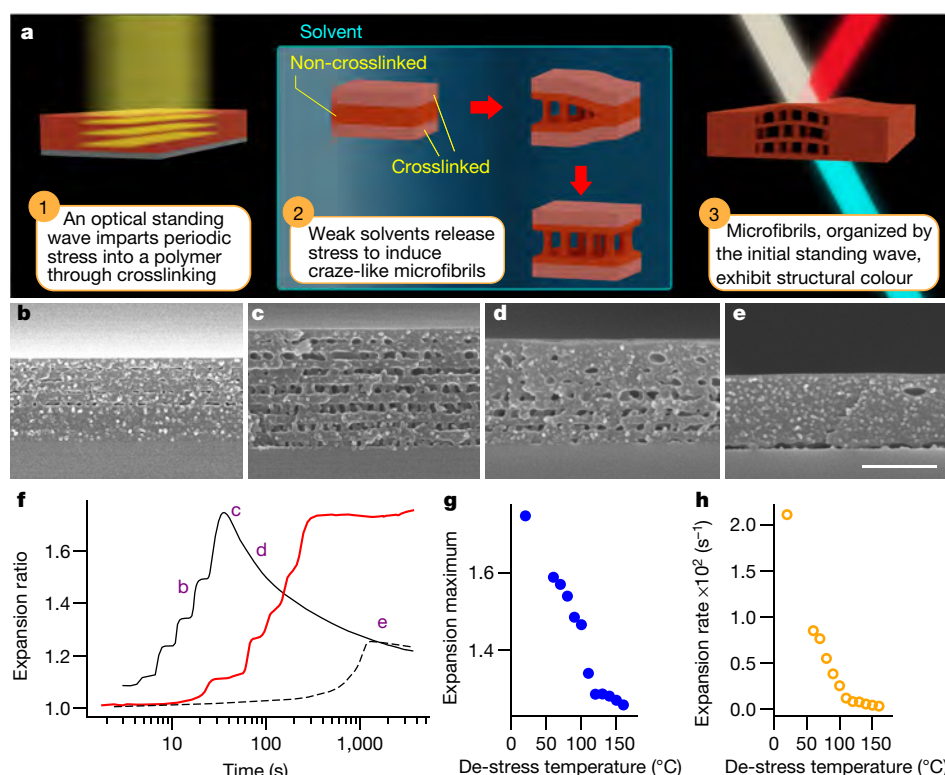


Fig. 1 | Organized stress microfibrillation phenomenon. **a**, Polymer films are crosslinked within a standing-wave pattern to impart internal stress fields. In a weak solvent the stress is released through the formation of microfibrils. The organization of microfibrils generate structural colour properties. **b–e**, Cross-sectional SEM images of 28-kDa polystyrene films crosslinked with wavelengths of 254 nm show the microfibril multilayer structure after different development times in acetic acid at 30 °C. **f**, The ratio between the expanded film thickness and the initial film thickness (the expansion ratio) of films during immersion in acetic acid. Solid black

and red lines refer to polystyrene films developed at 30 °C and 20 °C, respectively; the dashed line refers to a polystyrene film developed at 30 °C, which had first been de-stressed at 160 °C. Time points marked **b–e** in the plot refer to the SEM images in **b–e**. **g**, **h**, The peak expansion ratio (the expansion maximum) and the expansion rate (the expansion maximum divided by the time taken to reach the expansion maximum) for polystyrene films developed in 30 °C acetic acid after de-stressing at different temperatures. Scale bar for **b–e**, 500 nm.

of standing waves in forming the periodic layers by using polystyrene (PS)/PQ films of varying thickness crosslinked under light of wavelength 375 nm. Over a large range of film thicknesses, 55–1,050 nm, the number of porous layers increased at a regular interval (Fig. 2a), which matched the standing-wave interval of 375-nm light in polystyrene, 114 nm (Fig. 2b).

A range of LEDs of different wavelengths were used to crosslink the light-sensitive polystyrene films (of weight-averaged molecular weight 16 kDa). After development in acetic acid the resulting films had periodic structures and a Bragg peak whose position increased with increasing wavelength of the illumination light (Fig. 2c). The organized stress microfibrillation process was achieved with wavelengths of 285–405 nm in polystyrene. Changing the molecular weight of the polymer (to 160 kDa) results in a blueshift in the Bragg peak position (Fig. 2d, e). If the microfibrils can be associated with the production of crazing, this trend should be expected given that classic craze morphology forms more easily in lower-molecular-weight polymers^{12,13}.

The combined effect of increased illumination wavelength and the microfibril expansion increases the layer spacing of the periodic structure, pushing the Bragg peak into the visible light spectrum. The result is films with vivid structural colour^{14,15}. Even though the maximum illumination wavelength used was 405 nm, the films achieve colours that span the visible light spectrum, over 700 nm (Fig. 2d).

The organized stress microfibrillation process has been generalized to different molecular weights and illumination wavelengths within polystyrene films developed in acetic acid. To establish that this process can be generalized to other glassy polymers, we investigated a selection of frequently used polymers: bisphenol A polycarbonate, poly(methyl methacrylate) (PMMA), and polysulfone (PSF). Thin films of each

polymer were prepared, with BDABP added as the photoinitiator. Although our studies were initiated with PQ, it was found that BDABP was the more versatile photoinitiator in terms of its miscibility in other polymers, and it also had more efficient absorption at high wavelengths, leading to lower crosslinking times. The films were crosslinked under varying wavelengths of LED light and the porous layers were developed in solvents or solvent mixtures, such as acetic acid for polycarbonate, acetic acid and water for PMMA, and acetone and methanol for PSF. Microfibril multilayers could be formed in each of these polymers under different wavelengths of light (Fig. 3a–d).

Fracture in amorphous un-crosslinked glassy polymers is the result of the formation of crazes, which originate in points at which there is a triaxial stress state. If such stress is present, the crazes then grow and eventually give rise to fracture. In the presence of some substances, the onset of crazing occurs at lower stresses and their growth is faster. This phenomenon is called environmental stress crazing. In Hansen solubility theory, the solubility of a polymer/solvent pair can be characterized by a single value, the relative energy difference, where values of relative energy difference less than unity correspond to solvents that are likely to dissolve the polymer¹⁶. Crazing via environmental stress crazing is thought to occur when the polymer is exposed to solvents within a range of relative energy difference and molar volume values^{2,17}, where such values could be considered to represent the thermodynamic and kinetic conditions necessary for environmental stress crazing. Solvent systems with varying relative energy difference and molar volume were prepared and used to develop different polymer films. For each polymer, suitable solvent mixtures were identified that lie within a small region of the relative energy difference and molar volume parameter space (Extended Data Fig. 4). This similarity of solvent grouping to

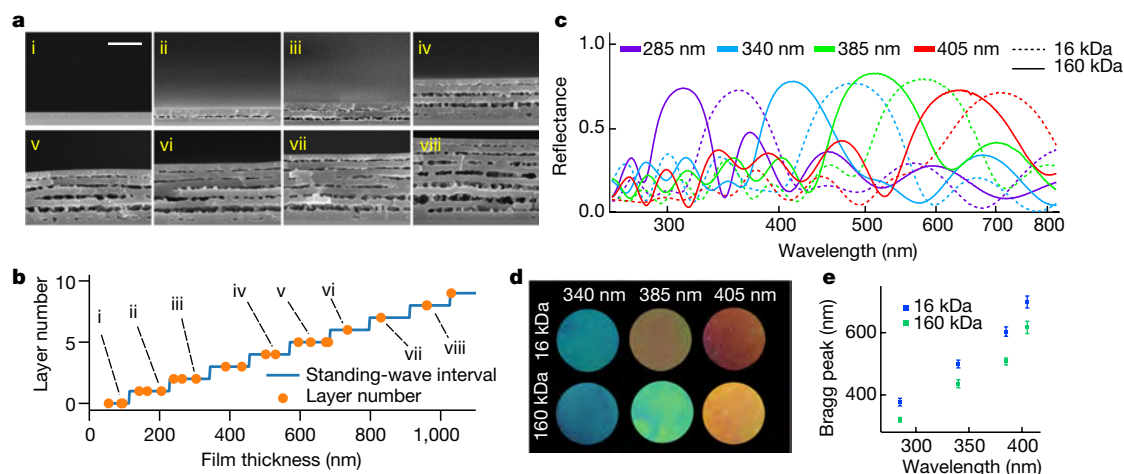


Fig. 2 | Structure from standing-wave interference. **a**, 35-kDa PS/PQ films of varying thickness were used to make organized microfibrillation structures. **b**, The number of porous layers in the films increases at regular intervals as the initial film thickness increases. This interval is equal to that of a standing wave within the as-cast films under 375-nm light, as indicated by the staircase function. **c**, The Bragg peak position of the microfibril structure for 16-kDa PS/PQ and 160-kDa PS/PQ increases

as the illumination wavelength increases. **d**, Films corresponding to the spectra in **c** demonstrate a range of structural colour from blue to red. **e**, Bragg peaks from lower-molecular-weight (16 kDa) polystyrene show that these films have greater expansion than the higher-molecular-weight (160 kDa) films. Error bars show one standard deviation with $n = 5$. Scale bar in **a**, 500 nm.

that observed in environmental stress crazing makes it plausible that the organized stress microfibrillation process shares characteristics with the general phenomenology of fracture in polymers. However, to establish that microfibrils are thus connected to crazing will require further characterization of the stress state in the films.

The organized stress microfibrillation process generates structural colour, and hence could be used to make images in polymer films. Using shadow masks, the illuminated region can be controlled to produce coloured images; by applying a different mask for each wavelength, multiple colour images can be created within the same film (Fig. 4a). For convenience, shadow masks can be generated by simple printing on traditional overhead transparency sheets. The printing process can also be applied to polymer films on transparent substrates (such as glass and polyester film). However, the standing-wave ratio is low for such systems (for example, 1.25 for polystyrene on polyethylene terephthalate (PET) at 375 nm) compared to the standing-wave ratio of polystyrene on Si (3.4 at 375 nm). The overall brightness of the structural colour on transparent substrates can be improved by placing the transparent substrate with the polymer film face down in contact with a silicon wafer so that irradiating light passes through the transparent substrate (Fig. 4b, c). However, uniform contact between the polymer film and silicon is difficult to achieve by this method. A comparison of the transmitted and reflected colour images on the transparent support

confirms that the colour thus generated is the result of bandgap reflections (Extended Data Fig. 5).

Masks made by conventional printing are limited by the resolution of the printer. Micrometre-level printing capability can be achieved using micro-LED illumination, thereby producing high-resolution images (Fig. 4d–l, Extended Data Fig. 6). Films with greater thickness can be used to minimize thin-film colour and to increase the colour contrast in the film (Fig. 4e). Using micro-LED illumination, feature sizes down to 2 μm can be achieved, as demonstrated using the US Airforce Resolution chart (Fig. 4j–l). Prints of periodic lines can achieve feature sizes of 1.8 μm or 14,000 dots per inch (Extended Data Fig. 6e).

Given the key role that thin-film stress plays in this microfibrillation process, it is unsurprising to see the influence of the residual stress from spin casting⁹. Therefore the image of the *Mona Lisa* painting, in which the film was prepared with a casting solvent of chloroform, as opposed to dichloromethane, shows larger variation in the colour (Fig. 4e, f). The colour change is also influenced by the feature size, as can be seen in the resolution chart (Fig. 4j–l). The feature-size colour effect was examined by printing images with pixel sizes of 20 μm and 10 μm . The resulting films have colours that differ according to the pixel size (Fig. 4g–i). Atomic force microscope surface probing shows that this effect is explained by the difference in expansion height between the two films; the 10- μm pixels did not expand as much as the 20- μm pixels,

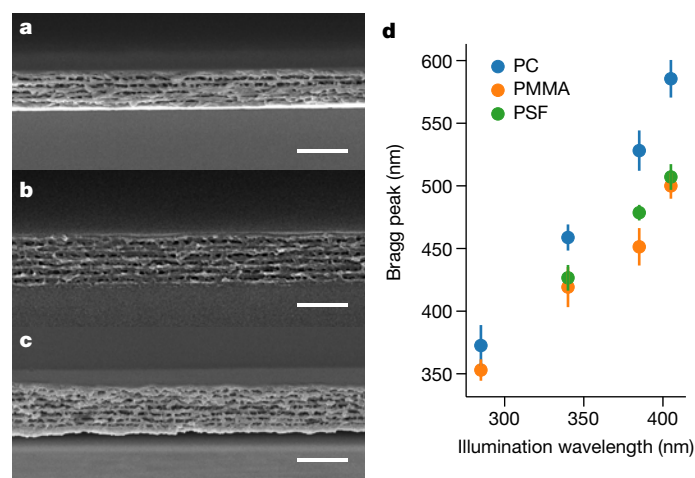


Fig. 3 | Organized stress microfibrillation in other polymers.

The microfibrillation structures can be formed in other polymers such as polycarbonate (in PC/BDABP) (**a**), poly(methyl methacrylate) (in PMMA/BDABP) (**b**), and polysulfone (in PSF/BDABP) (**c**). **d**, Bragg peak position versus illumination wavelength for polycarbonate, PMMA and PSF. Error bars show one standard deviation with $n = 5$. Scale bar for **a–c**, 1 μm .

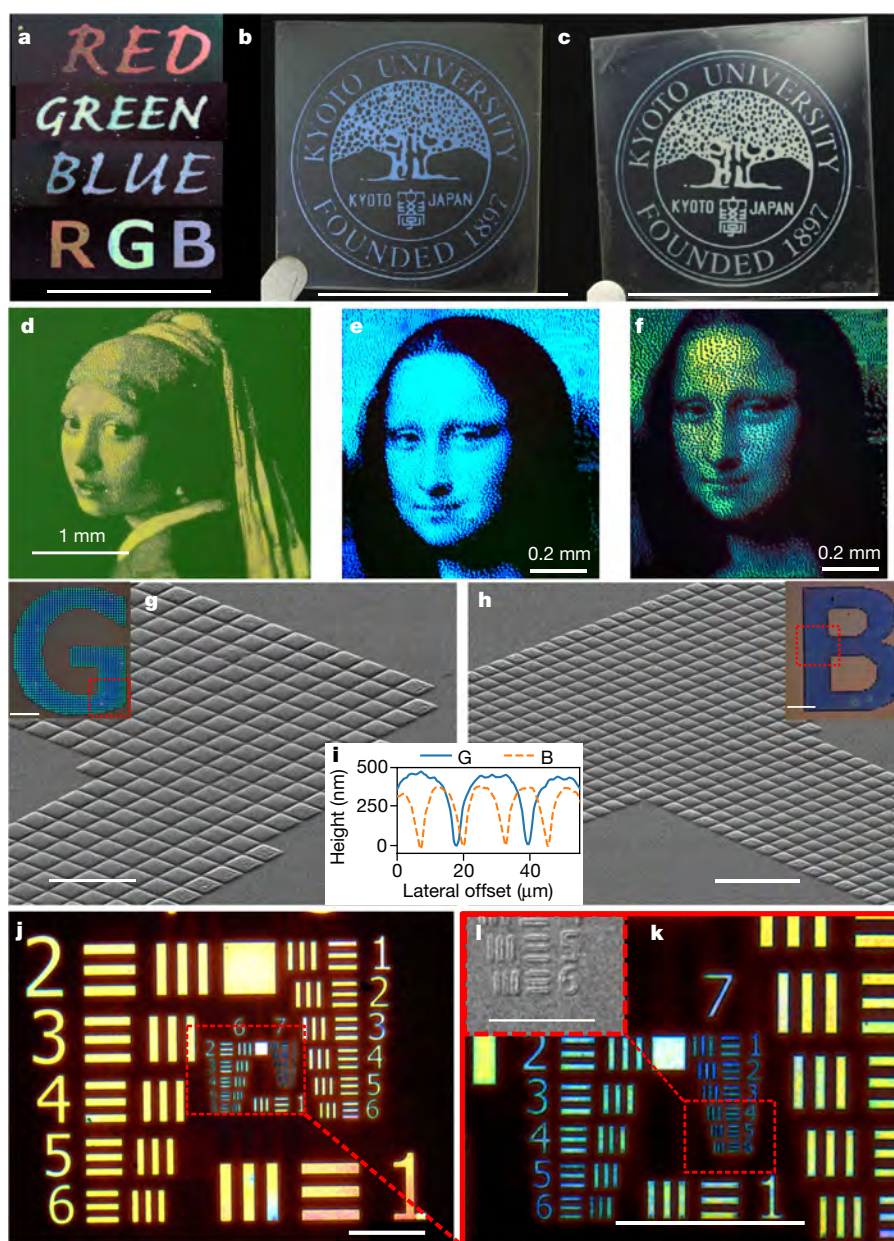


Fig. 4 | Applications of organized stress microfibrillation method.

a, Overhead projector shadow masks were used to generate the red, green and blue text in 35-kDa PS/PQ using 405-nm, 385-nm and 340-nm light, respectively. Thin polydimethyl siloxane (PDMS) sheets were used to isolate structural colour from thin-film interference colour. **b, c**, Structural colour images printed in 192-kDa PS/BDABP on transparent PET substrates with 340-nm light (**b**) and 385-nm light (**c**). **d**, Image of the *Girl with a Pearl Earring* painting created via micro-LED illumination of 35-kDa PS/PQ film. **e, f**, Images of the *Mona Lisa* painting created by micro-LED illumination of 192-kDa PS/BDABP film spin-cast with dichloromethane (**e**) and chloroform (**f**). **g, h**, SEM

images of the letters G and B created with 20- μ m and 10- μ m micro-LED pixels, respectively, in 192-kDa PS/BDABP. The insets show microscopic images. **i**, Atomic force microscopy indicates that the 20- μ m pixels have larger expansion than the 10- μ m pixels. **j, k**, Optical microscope images of the US Airforce resolution chart in 192-kDa PS/BDABP demonstrates the resolution range achievable from micro-LED printing of organized microfibrillation structures. The region shown in **k** is indicated by the dotted outline in **j**. Similarly, the inset (**l**) shows an SEM image of the outlined region in **k**. Scale bars: **a**, 1 cm; **b** and **c**, 3 cm; **g** and **h**, 50 μ m (inset 200 μ m); and **j** and **k**, 200 μ m; **l**, 50 μ m.

probably because the expansion is pinned by the edges of the pixel. The colour-feature size dependence provides an additional way to introduce colour into films illuminated with monochromatic light.

Chemical transformation of nanoparticles, micelles and photoresists within optical interference fields is the basis of holography technology, with the aim of developing structural colour and microstructural devices^{4,14,18–24}. There has been a recent shift towards using fracture proactively as a mechanism to create the microstructure for such devices^{25–29}. Our design principle uses optically generated stress fields to develop organized, tunable microstructures within polymers.

Online content

Any Methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available in the online version of the paper at <https://doi.org/10.1038/s41586-019-1299-8>.

Received: 30 July 2018; Accepted: 30 April 2019;

Published online 19 June 2019.

1. Kambour, R. P. Review of crazing and fracture in thermoplastics. *J. Polym. Sci. D* **7**, 1–154 (1973).
2. Robeson, L. M. Environmental stress cracking: a review. *Polym. Eng. Sci.* **53**, 453–467 (2013).

3. Ward, A. L., Lu, X., Huang, Y. & Brown, N. The mechanism of slow crack growth in polyethylene by an environmental stress cracking agent. *Polymer* **32**, 2172–2178 (1991).
4. Connes, P. Silver salts and standing waves: the history of interference colour photography. *J. Opt.* **18**, 147–166 (1987).
5. Kumar, K. et al. Printing colour at the optical diffraction limit. *Nat. Nanotechnol.* **7**, 557–561 (2012).
6. Carroll, K. M. et al. Fabricating nanoscale chemical gradients with thermochemical nanolithography. *Langmuir* **29**, 8675–8682 (2013).
7. Sakoda, K. *Optical Properties of Photonic Crystals* Vol. 80 (Springer, 2004).
8. Born, M. & Wolf, E. *Principles of Optics: Electromagnetic Theory of Propagation, Interference and Diffraction of Light* (Elsevier, 2013).
9. Chung, J. Y., Chastek, T. Q., Fasolka, M. J., Ro, H. W. & Stafford, C. M. Quantifying residual stress in nanoscale thin polymer films via surface wrinkling. *ACS Nano* **3**, 844–852 (2009).
10. Kang, C. et al. Full color stop bands in hybrid organic/inorganic block copolymer photonic gels by swelling–freezing. *J. Am. Chem. Soc.* **131**, 7538–7539 (2009).
11. Li, T., Zhou, C. & Jiang, M. UV absorption spectra of polystyrene. *Polym. Bull.* **25**, 211–216 (1991).
12. Bray, J. C. & Hopfenberg, H. B. The effect of polymer molecular weight on the solvent crazing of polystyrene. *J. Polym. Sci. B* **7**, 679–684 (1969).
13. Donald, A. M. & Kramer, E. J. Effect of molecular entanglements on craze microstructure in glassy polymers. *J. Polym. Sci. Polym. Phys. Ed.* **20**, 899–909 (1982).
14. Kolbe, M. et al. Mimicking the colourful wing scale structure of the *Papilio blumei* butterfly. *Nat. Nanotechnol.* **5**, 511–515 (2010).
15. Chung, W.-J. et al. Biomimetic self-templating supramolecular structures. *Nature* **478**, 364–368 (2011).
16. Hansen, C. M. 50 years with solubility parameters—past and future. *Prog. Org. Coat.* **51**, 77–84 (2004).
17. Hansen, C. M. & Just, L. Prediction of environmental stress cracking in plastics with Hansen solubility parameters. *Ind. Eng. Chem. Res.* **40**, 21–25 (2001).
18. Nadal, E. et al. Plasmon-enhanced diffraction in nanoparticle gratings fabricated by in situ photo-reduction of gold chloride doped polymer thin films by laser interference patterning. *J. Mater. Chem. C* **5**, 3553–3560 (2017).
19. Matsubayashi, A., Fukunaga, K., Tsuji, T., Ataka, K. & Ohsaki, H. Multilayered ordering of the metal nanoparticles in polymer thin films under photoirradiation. *Langmuir* **27**, 733–740 (2011).
20. Smirnov, J. R. C. et al. Adaptable ultraviolet reflecting polymeric multilayer coatings of high refractive index contrast. *Adv. Opt. Mater.* **3**, 1633–1639 (2015).
21. Park, S.-G. & Yang, S.-M. Multicolor patterning using holographic woodpile photonic crystals at visible wavelengths. *Nanoscale* **5**, 4110 (2013).
22. Yuan, L. & Herman, P. R. Laser scanning holographic lithography for flexible 3D fabrication of multi-scale integrated nano-structures and optical biosensors. *Sci. Rep.* **6**, 22294 (2016).
23. Suzuki, N., Tomita, Y. & Kojima, T. Holographic recording in TiO₂ nanoparticle-dispersed methacrylate photopolymer films. *Appl. Phys. Lett.* **81**, 4121–4123 (2002).
24. Campbell, M., Sharp, D. N., Harrison, M. T., Denning, R. G. & Turberfield, A. J. Fabrication of photonic crystals for the visible spectrum by holographic lithography. *Nature* **404**, 53–56 (2000).
25. Nam, K. H., Park, I. H. & Ko, S. H. Patterning by controlled cracking. *Nature* **485**, 221–224 (2012).
26. Choi, Y. W. et al. Ultra-sensitive pressure sensor based on guided straight mechanical cracks. *Sci. Rep.* **7**, 40116 (2017).
27. Dubois, V., Niklaus, F. & Stemme, G. Design and fabrication of crack-junctions. *Microsyst. Nanoeng.* **3**, 17042 (2017).
28. Yang, C. et al. All-solution-processed, scalable, self-cracking Ag network transparent conductor. *Phys. Status Solidi* **215**, 1700504 (2018).
29. Kim, M., Ha, D. & Kim, T. Cracking-assisted photolithography for mixed-scale patterning and nanofluidic applications. *Nat. Commun.* **6**, 6247 (2015).

Acknowledgements We thank D. Hirayama for assistance with sample preparation and the support of K. Kuroda, head of the JST-Presto Sakigake programme. We thank the Analysis Centre at iCeMS, KUIAS, Kyoto University for access to their SEM and confocal microscope. We thank the Nanohub at Kyoto University for access to their clean room facilities and micro-LED. This work was supported by JST-PRESTO (JPMJPR1417) and the central facilities are supported by the World Premier International Research Initiative (WPI), MEXT, Japan.

Reviewer information Nature thanks Seung Hwan Ko, Shin-Hyun Kim and Marta Rink for their contribution to the peer review of this work.

Author contributions M.M.I. and E.S. initiated, managed and planned the overall project. M.M.I. established the printing protocols. A.H.G. developed the microscopy elements of this work. D.Q., D. Yamamoto, D. Yamaguchi and H.J. all performed the experiments. All authors contributed to data analysis and manuscript refinement and preparation.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-019-1299-8>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-019-1299-8>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to M.M.I. or E.S.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

METHODS

Polymers. Specialty-grade polystyrenes (with weight-averaged molecular weight, M_w 16 kDa, 28 kDa and 160 kDa, polydispersity < 1.1 as determined by manufacturer) were obtained from Polymer Source. Commercial-grade polystyrenes (M_w 35 kDa and 192 kDa, polydispersity > 1.5 as measured using a gel permeation chromatograph (GPC), bisphenol A polycarbonate (M_w 45 kDa, polydispersity > 1.5 as measured with a GPC), poly(methyl methacrylate) (PMMA, M_w 120 kDa, polydispersity > 1.5 as measured with a GPC), and polysulfone (PSF, M_w 35 kDa, polydispersity > 1.5 as measured with a GPC) were obtained from Sigma-Aldrich. A photoinitiator 9,10-phenanthrenequinone (PQ) and 4,4'-bis-(diethylamino)-benzophenone (BDABP) were obtained from Tokyo Chemical Industry and Sigma-Aldrich, respectively. Solvents *n*-hexane, toluene, dichloromethane, chloroform, methanol, ethanol, tetrahydrofuran, acetic acid, ethyl acetate and butyl acetate were obtained from Nacalai Tesque. Solvents 1-propanol, 2-propanol, 1-butanol, 1-chloropentane and acetonitrile, and aluminium oxide powder (Brockmann I) were obtained from Sigma-Aldrich. PDMS was obtained from Dow Corning. Deionized water was produced in the laboratory using a Milli-Q Type 1 Ultrapure Water System (Merck-Millipore).

Purification of polymers. Commercial-grade polystyrene, polycarbonate, PMMA and PSF were purified according to the following protocol: 3 g of polymer was dissolved in 50 ml solvent (toluene for polystyrene, chloroform for polycarbonate, PMMA and PSF) and sonicated for 30 min at 50 °C. The solution was filtered through a 0.2- μ m-pore PTFE membrane syringe filter (md Membrane Technologies) with aluminium oxide powder packed into the syringe. The filtrate was mixed with 200 ml deionized water in a flask and vigorously shaken for 1 min. The mixture was left to separate for 20 min after which the water was drained. The water mixing/draining steps were repeated another four times. 200 ml of methanol was added dropwise to the solution to reprecipitate the polymer. The precipitate was filtered (Whatman filter paper, grade 41) for collection. The purified polymer was dried in a vacuum oven for two days at 50 °C and stored in a desiccator.

Formation of organized microfibrillation structures. Three steps are performed in sequence: spin-casting, crosslinking, and solvent development³⁰.

Spin-casting. The prepared polymer/photoinitiator solutions in either dichloromethane or chloroform were sonicated for 20 min and filtered through a 0.2- μ m-pore PTFE membrane syringe filter. Afterwards, polymer solutions were spun-cast on a clean substrate using a spincoater (MS-A100, Mikasa). Spin casting speeds and concentrations (0.5–5 weight per cent polymer in solvent) were adjusted to obtain desired film thicknesses. Photoinitiator content was 6 and 12 weight per cent in the polymer when BDABP and PQ were used, respectively. Silicon wafer (single side polished), $< 100 \rangle$, N-type (Matsuzaki Seisakusyo, FC-100), cover glass (Muto Pure Chemicals), and PET sheets (Lumirror T60, 250- μ m-thick, Toray Industries) were the three typical substrates employed in this study.

Crosslinking. For large-area crosslinking, the spin-cast polymer films were illuminated in crosslinking ovens. The CX-2000 (UVP) oven was used for 254 nm illumination. Custom ovens were constructed for LED light sources (Thorlabs; wavelengths 300, 340, 375, 385, 395 and 405 nm; see Extended Data Fig. 1). For each wavelength, six LEDs were assembled into an array (2 columns \times 3 rows) and fixed to the ceiling of a custom enclosure (height \times width \times length = 10 cm \times 10 cm \times 18 cm). A USB fan (Evercool, UFAN-12) was operated on top of each box to maintain the custom ovens at room temperature during irradiation.

The placement of polymer films during crosslinking depends on the nature of the substrate. Polymer films on a reflective substrate (for example, silicon wafer) were placed flat, face-up towards the light source, while polymer films on a transparent substrate (for example, cover glass or PET sheet) were placed face-down on a mirrored surface (for example, silicon wafer) with irradiating light passing through the transparent substrate.

To make large-scale images in the films, shadow masks were placed on top of the films during the illumination to crosslink according to the mask pattern. The shadow mask can be custom-designed with graphics software and printed onto transparent overhead projector sheets (Folex FX-471P).

For microscale pattern printing, polymer films were exposed to a micro-LED light source (Maskless Lithography Tool D-light DL-1000GS/KCH, Nano System Solutions). The D-light DL1000 is combined with a digital mirror device (Texas Instruments) for controlling the projection area. The digital mirror device is an electro-mechanical system consisting of 1,024 \times 768 micro-mirrors. The on/off state of each mirror can be controlled individually. The individual mirrors reduce and project light onto 1 μ m \times 1 μ m regions through a lens, which compose to produce an exposure area of 1,024 μ m \times 768 μ m as a single frame. Larger images can be stitched together from multiple frames. An LED with wavelength of 405 nm and intensity of 20 mW cm⁻² was used as the light source of the D-light DL-1000. Typical dose amounts varied from 40–160 J cm⁻² and 500 J cm⁻² for BDABP- and PQ-based thin films respectively. Patterns for micro printing were designed with computer-aided design (CAD)

software (L-Edit) (<https://www.mentor.com/tannereda/l-edit>) and transferred to the micro-LED lithography system as the data input. An outline of the CAD image conversion steps can be found in Extended Data Fig. 6.

Development. The crosslinked polymer films were immersed in a suitable weak solvent. Acetic acid was the typical development solvent. Other solvents and solvent mixtures are shown in Extended Data Fig. 4. Solvent development was performed at room temperature (20 °C) unless otherwise stated. The development process was completed based on the observation of colour changes in the polymer film. Upon completion, developed films were removed and dried using an air blower.

Characterization. An SEM (JEOL JSM-7500F) was used to observe polymer film structures. For cross-sectional observation the samples were fractured while submerged in liquid nitrogen. Each sample was coated with 5 nm of osmium via sputter coating (Filgen Osmium Plasma Coater OPC60A). The topography of polymer films was investigated with an atomic force microscope (NanoWizard III, JPK Instruments) in non-contact mode.

A digital single-lens reflex camera (EOS Kiss X5, Canon) with macro-lens (60-mm EFS, Canon) was employed to take photographs of samples. Before taking photographs the camera was white-balanced using an 18% neutral grey card. When we wanted to eliminate the thin-film interference effect and isolate the structural colour, 2-mm-thick PDMS films were placed on top of the films. PDMS films were prepared with a Sylgard 184 kit by the following steps: (1) the base component was mixed well with the crosslinking agent in a ratio of 10/1 (by weight); (2) the solution was degassed in vacuum; and (3) the solution was poured into a Petri dish and allowed to set overnight at 50 °C into a flexible film.

An upright optical microscope (Axioscope A1 MAT, Carl Zeiss) was used to obtain images of the micro-LED printed polymer films. Samples were illuminated from above with a 100-W halogen bulb light source.

Reflectance spectra were acquired using an ultraviolet–visible spectrometer (MCPD-3700, Otsuka Electronics) with a 210–820-nm light source (MC-2530, Otsuka Electronics). Optical analysis software was used to determine the film thickness using the Cauchy equation model⁸. All reflectance spectra were normalized by the substrate (bare silicon wafer).

A GPC (Shimadzu UFLC system) was used to analyse the molecular weight distribution of the polymers. Analysed polymers were dissolved in tetrahydrofuran to a concentration of 0.2 weight per cent.

In situ spectrum measurement. The effects of stress are of interest in the in situ experiments. To remove pre-existing stress in the films, from spin-cast, polystyrene films (M_w 28 kDa, Polymer Source) were pre-annealed in a custom vacuum hotplate at 190 °C for 2 h and cooled down to room temperature. After crosslinking under 254-nm light the films were thermally de-stressed by post-annealing in the custom vacuum hotplate under varying temperatures for one hour and allowed to cool down to room temperature. One further sample was annealed for much longer (48 h at 160 °C) so that we could observe the limits of thermally de-stressing. The development solvent, acetic acid, was kept at a stable temperature (30 °C or 20 °C) in a glass dish using a hotplate (AS ONE, CHPS-250DN). The reflectance spectra of the polymer films were measured using an ultraviolet–visible spectrometer (CX-2000, Otsuka) in situ during its immersion in acetic acid under continuous mode. The measurement interval was 0.6 s for the initial 10 min and was switched to a 3-s interval for the remaining time. The change in film thickness during development was calculated through the analysis of the in situ spectrum.

Analysis of in situ spectra. The temporal development of the in situ reflection spectrum was modelled and fitted with a two-layer structure of acetic acid and swelling polystyrene. The model is described as follows⁸:

$$I_{\text{power}} = R_{\text{total}} R_{\text{total}}^* \quad (1)$$

$$R_{\text{total}} = (r_{\text{airAA}} + R_f) / (1 - r_{\text{AAair}} R_f) \quad (2)$$

$$R_f = (r_{\text{AAf}} + r_{\text{isi}} \exp(-2i\theta)) / (1 - r_{\text{fAA}} r_{\text{isi}} \exp(-2i\theta)) \quad (3)$$

where an asterisk denotes complex conjugation. I_{power} is the power reflection coefficient, R_{total} is the total reflection coefficient of the structure, R_f is the reflection coefficient from the polystyrene/acetic acid interface and θ is the transmission phase change. r_{12} denotes the reflection coefficient at the interfaces between materials 1 and 2, specifically:

$$r_{\text{fAA}} = (n_f - n_{\text{AA}}) / (n_f + n_{\text{AA}})$$

$$r_{\text{AAf}} = -r_{\text{fAA}}$$

$$r_{\text{AAair}} = (n_{\text{AA}} - 1) / (1 + n_{\text{AA}})$$

$$r_{\text{airAA}} = -r_{\text{AAair}}$$

$$r_{\text{fsi}} = (n_f - n_{\text{Si}} + i\kappa_{\text{Si}}) / (n_f + n_{\text{Si}} - i\kappa_{\text{Si}})$$

where the subscripts refer to the materials (air, film, acetic acid and silicon). The refractive index of the expanded film, n_f , is a combination of the polystyrene refractive index and that of acetic acid. It is calculated by the Lorentz–Lorenz equation with the fraction of acetic acid, φ , as the fitting parameter³¹:

$$\frac{n_f^2 - 1}{n_f^2 + 2} = \frac{n_{\text{AA}}^2 - 1}{n_{\text{AA}}^2 + 2} \varphi + \frac{n_{\text{PS}}^2 - 1}{n_{\text{PS}}^2 + 2} (1 - \varphi)$$

where n_{PS} , the refractive index of the polymer (polystyrene), is calculated via the Sellmeier model³², and the refractive index of acetic acid n_{AA} is obtained from experimental values³³. The refractive index of silicon, n_{Si} , and its extinction coefficient, κ_{Si} , are obtained from experimental values³⁴. The expansion ratio is calculated as $1/(1-\varphi)$.

Standing-wave finite difference time domain simulation. The standing-wave behaviour within the films at varying thickness was simulated using the finite difference time domain method³⁵. The open-source software package MEEP was used to perform the simulations³⁶. A one-dimensional domain was used with an incoming monochromatic plane wave (385 nm). Polystyrene was modelled as a layer with thickness varied over a number of simulations; the refractive index of the layer was derived using the Sellmeier model. The silicon substrate was modelled as a layer containing a perfectly matched layer to absorb incoming waves, allowing the silicon to act as a semi-infinite layer.

Standing-wave ratio. The standing-wave ratio is a measure of the contrast of a standing wave and is defined as the ratio between the anti-node peak amplitude and node peak amplitude. Using the Fresnel equation to determine the reflection coefficient, r , the standing-wave ratio can be calculated as $\frac{1+|r|}{1-|r|}$.

Photonic multilayer films. For dried, acetic-acid-free samples, the transfer matrix method is the most analytical way to fit the measured spectra in a physically meaningful way, because we can assume a multilayer model of alternating high- and low-density layers, where the low-density layer corresponds to the fibrillated structure. It allows a relatively accurate approximation of both porosity and layer sizes. The transfer matrix method is applied in a coarse-grained approach to determine the overall film thickness, and its expansion, during immersion in acetic acid.

The developed film structure was modelled as a one-dimensional photonic multilayer. The transfer matrix method was used to model the structure^{8,37}. We assumed that each of the solid layers had the same thickness and that each of the porous layers had the same thickness. The refractive index of the porous layer, n_p , was approximated as the volume-weighted average of the refractive indices of polystyrene and air⁸:

$$n_p = p n_{\text{air}} + (1-p) n_{\text{PS}}$$

where n_{air} is the refractive index of air, n_{PS} is the refractive index of polystyrene and p is the volume fraction of air (the porosity). Model parameters for layer thickness were obtained from SEM cross-sections and were fitted according to the least-squares method where the spectra were fitted in order to deduce the porosity of the films.

Polymer solubility analysis. According to Hansen solubility parameter theory^{2,38}, the solubility of a polymer depends on: the energy from dispersion forces between molecules (δ_D), the energy from dipolar intermolecular forces between molecules (δ_P), and the energy from hydrogen bonds between molecules (δ_H). These three parameters form the three-dimensional Hansen solubility space. A solvent is represented as a point in Hansen space and a polymer is described by a sphere with radius R_0 . The relative energy difference is defined as R_a/R_0 , where $R_a^2 = 4(\delta_{D1}-\delta_{D2})^2 + (\delta_{P1}-\delta_{P2})^2 + (\delta_{H1}-\delta_{H2})^2$ and subscripts 1 and 2 denote a solvent and a polymer, respectively.

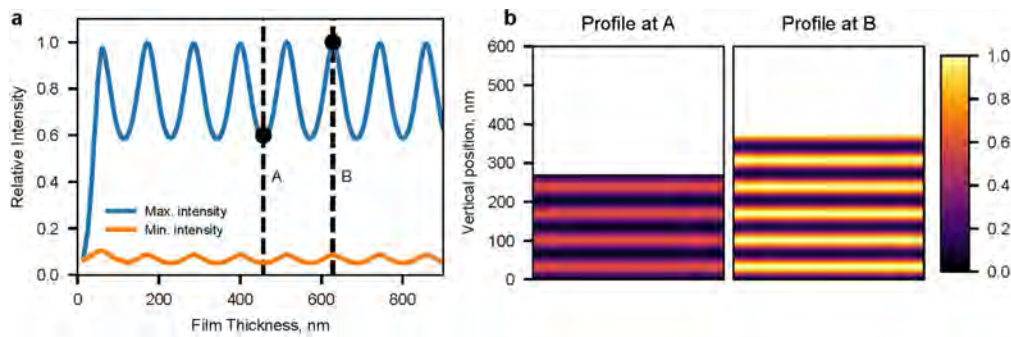
Code availability

The code used for optical analysis in this study is available from the corresponding authors upon reasonable request.

Data availability

The data that supports the findings of this study are available from the corresponding authors upon reasonable request.

- Thurn-Albrecht, T. et al. Nanoscopic templates from oriented block copolymer films. *Adv. Mater.* **12**, 787–791 (2000).
- Hori, K., Matsuno, H. & Tanaka, K. Sorption kinetics of methanol in thin poly(methyl methacrylate) films studied by optical reflectivity. *Soft Matter* **7**, 10319 (2011).
- Nikolov, I. D. & Ivanov, C. D. Optical plastic refractive measurements in the visible and the near-infrared regions. *Appl. Opt.* **39**, 2067–2070 (2000).
- El-Kashef, H. The necessary requirements imposed on polar dielectric laser dye solvents. *Phys. B* **279**, 295–301 (2000).
- Aspnes, D. E. & Studna, A. A. Dielectric functions and optical parameters of Si, Ge, GaP, GaAs, GaSb, InP, InAs, and InSb from 1.5 to 6.0 eV. *Phys. Rev. B* **27**, 985–1009 (1983).
- Taflove, A. & Hagness, S. C. *Computational Electrodynamics: The Finite-Difference Time-Domain Method* (Artech, 2005).
- Oskooi, A. F. et al. Meep: A flexible free-software package for electromagnetic simulations by the FDTD method. *Comput. Phys. Commun.* **181**, 687–702 (2010).
- Li, Z.-Y. & Lin, L.-L. Photonic band structures solved by a plane-wave-based transfer-matrix method. *Phys. Rev. E* **67**, 046607 (2003).
- Hansen, C. M. On predicting environmental stress cracking in polymers. *Polym. Degrad. Stab.* **77**, 43–53 (2002).
- Akay, C., Parrein, P. & Rolland, J. P. Estimation of longitudinal resolution in optical coherence imaging. *Appl. Opt.* **41**, 5256–5262 (2002).
- Wu, D. Y., Meure, S. & Solomon, D. Self-healing polymeric materials: a review of recent developments. *Prog. Polym. Sci.* **33**, 479–522 (2008).



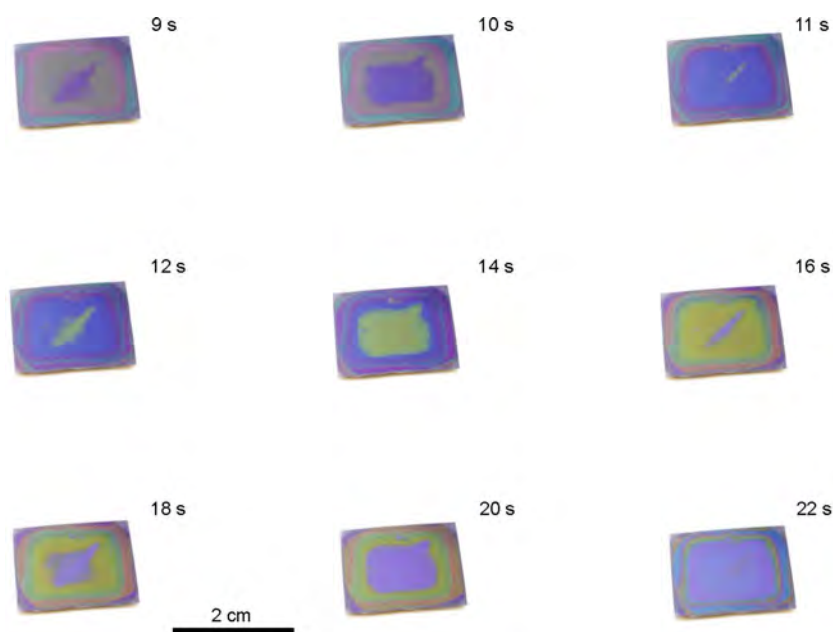
LED central wavelength (nm)	Full width half maximum, $\Delta\lambda$ (nm)	Coherence Length, (μm)
285	13	1.61
300	20	1.17
340	11	2.79
375	9	4.19
385	12	3.33
395	16	2.63
405	12	3.70

Extended Data Fig. 1 | Standing-wave modelling and light coherence. **a**, Reflecting light within the thin film interferes to form standing waves. As the thickness of the film increases, the minimum intensity remains close to zero while the maximum intensity of the standing wave oscillates within the film, as shown by finite difference time domain simulations of polystyrene on silicon. **b**, Profiles of standing waves in films of different thickness. The colour shows the light intensity relative to the maximum light intensity that can be achieved in this system (the colour scale shows relative light intensity, unitless). At any thickness, for example those indicated by A and B in **a**, standing-wave interference occurs, although the anti-node intensity can vary by a factor of up to 0.6 in this polystyrene

system. Fluctuations in thickness of the thin films therefore do not prevent crosslinking from occurring as long as the applied dose is high enough. The table below outlines the coherence length of the LED light sources used. Approximating Gaussian light sources, the coherence length is

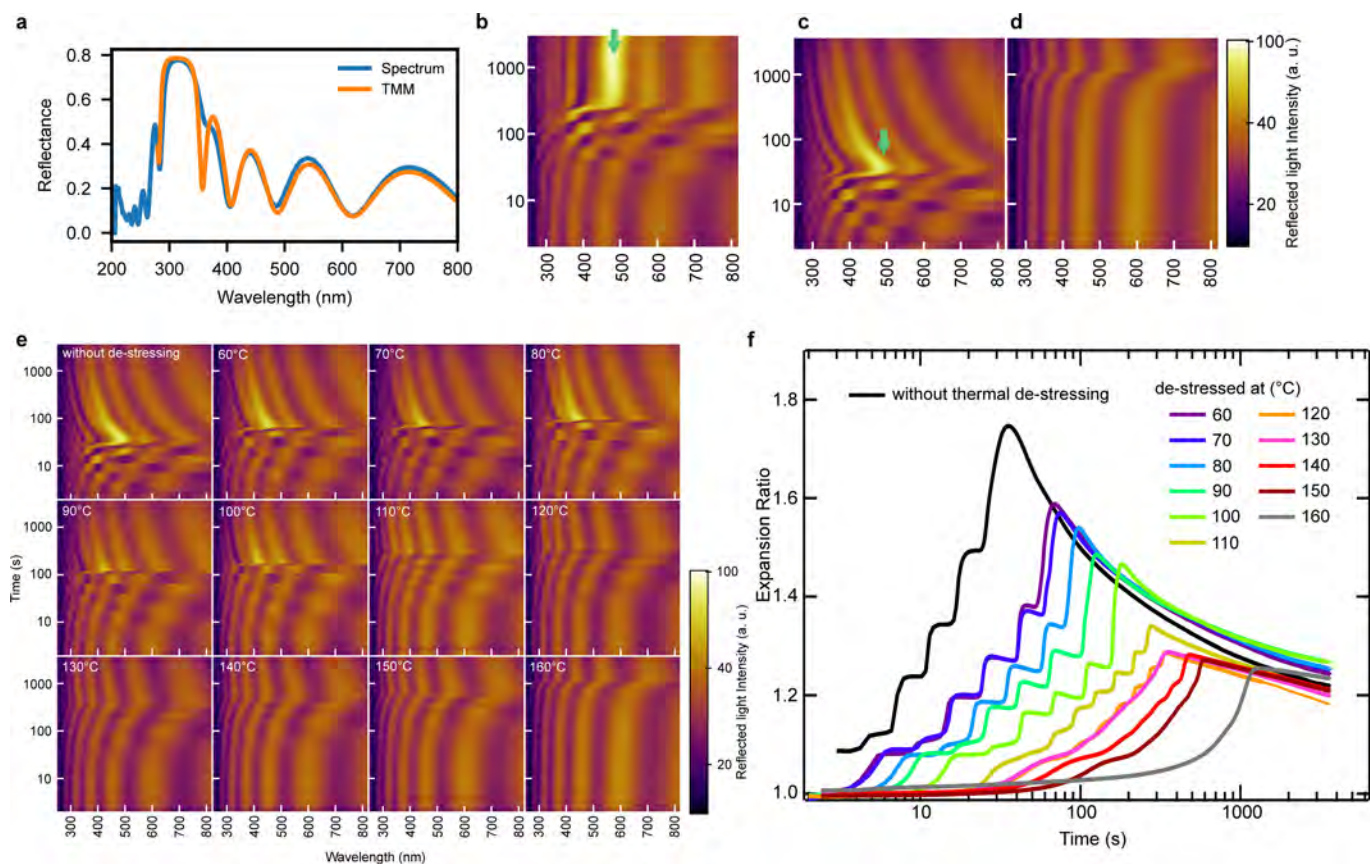
$$l_c = \sqrt{\frac{2 \ln(2)}{\pi}} \frac{\lambda_0^2}{n \Delta \lambda_0}$$

where λ_0 is the free space wavelength of the light source, $\Delta \lambda_0$ is the spectral width (full-width at half-maximum, FWHM), and n is the refractive index of the medium³⁹. The polymer films in this study had thicknesses of less than 1 μm so that coherent interference could be achieved with light sources that have coherence lengths on the micrometre scale.



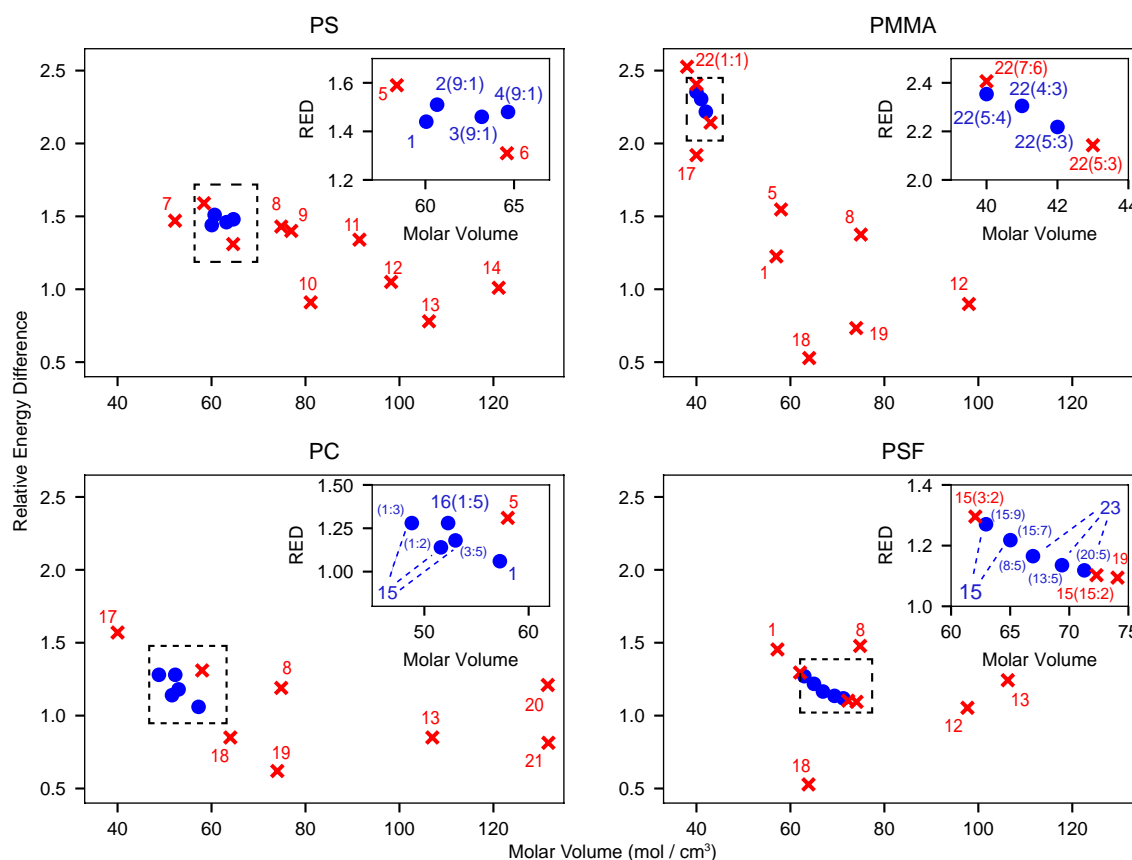
Extended Data Fig. 2 | Colour change during layer formation in acetic acid. Crosslinked polystyrene film (28 kDa) is submerged in acetic acid and porous layers are formed within one minute. Images show snapshots of the stepwise colour change that occurs during the formation of the layers. The time after submersion is indicated in each image. Images were captured using a digital single-lens reflex camera with a macro-lens (EOS

Kiss X5, Canon and 60-mm EFS, Canon); see Supplementary Video 1. We found that the large-scale colour change can begin from any location on the film including the centre (as shown), edges and corners, depending on the local film condition. The final structure and in-plane layer formation dynamics as discussed within this paper are independent of these large-scale colour change effects.



Extended Data Fig. 3 | Spectroscopic analysis of the development of films. **a**, Reflectance spectrum of the developed polystyrene (28 kDa) films and a model photonic multilayer film spectrum produced using the transfer matrix method. **b–d**, Real-time spectroscopy of the immersion process in acetic acid at various conditions. **b**, Polystyrene film crosslinked and developed in an acetic acid bath (20 °C). The film initially undergoes a series of sharp changes in spectra, indicating changes in film thickness, and then after formation of a Bragg peak (green arrows), the spectrum persists. **c**, Polystyrene film crosslinked and developed in an acetic acid bath (30 °C). Unlike in **b**, the final Bragg peak decays. **d**, Polystyrene film crosslinked, and then annealed at 160 °C for 2 days and developed in an acetic acid bath (30 °C). The heat map shows a smooth change in spectra, indicating a gradual increase towards a final swollen film without

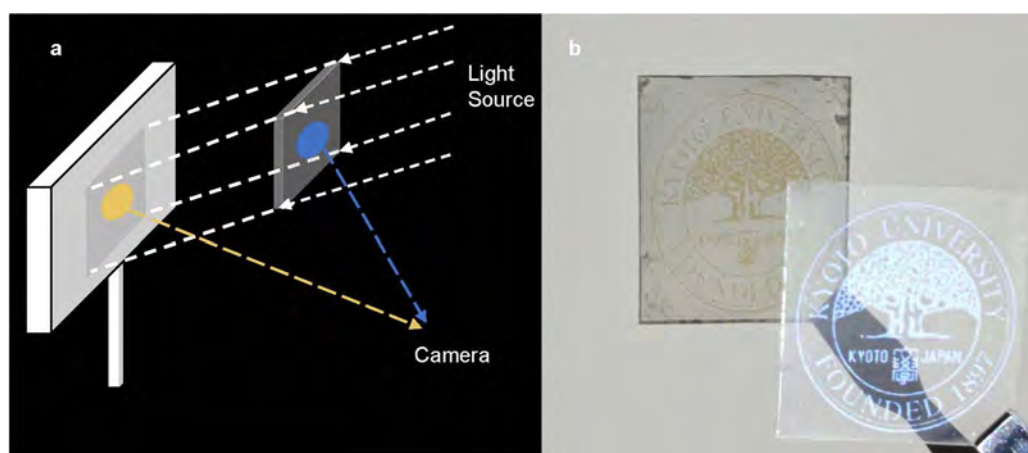
microfibril formation. **e**, After crosslinking, films were post-annealed at a range of temperatures (60–160 °C) to remove stresses in the films (thermal de-stressing). Clear stepwise changes in the spectrum are seen in films that were thermally de-stressed below the glass transition temperature (approximately 110 °C). After thermal de-stressing at higher temperatures the stepwise expansion effect is reduced and the films simply expand smoothly. **f**, The film thickness expansion over time is derived from real-time spectra where 1.0 corresponds to the initial film thickness. After microstructure expansion finishes the films relax to equilibrium thickness. Without any stress or external force, crazes in glassy polymers undergo collapse under certain conditions⁴⁰, as seen by the decay of the expansion ratio.



No.	Solvent	No.	Solvent
1	Acetic acid	13	Toluene
2	Ethanol / THF	14	1-Chloropentane
3	Ethanol / Toluene	15	Acetone / Methanol
4	Ethanol / 1-Chloropentane	16	Toluene / Methanol
5	Ethanol	17	Methanol
6	Acetonitrile / 1-Propanol	18	Dichloromethane
7	Acetonitrile	19	Acetone
8	1-Propanol	20	Hexane
9	2-Propanol	21	Butyl Acetate
10	THF	22	Acetic Acid / Water
11	Butanol	23	Acetone / Acetic Acid
12	Ethyl Acetate		

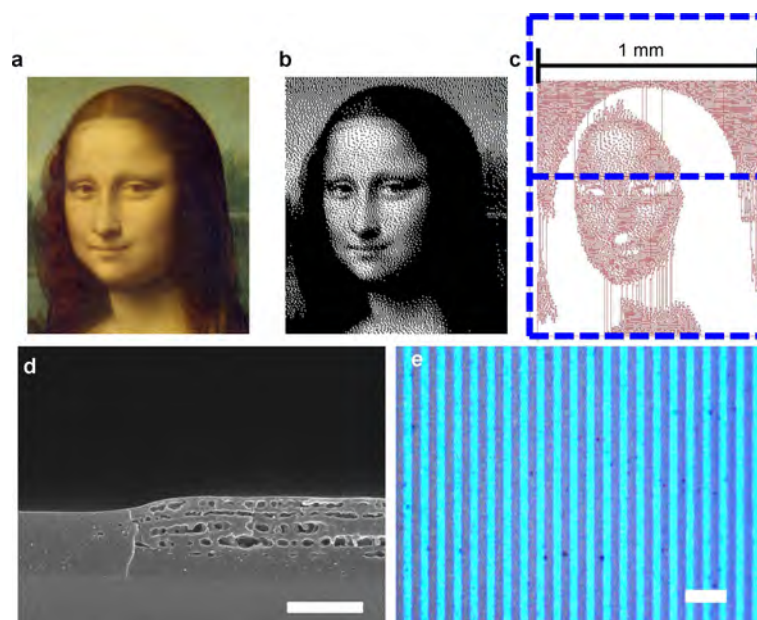
Extended Data Fig. 4 | Hansen parameter plots. Specific solvent conditions associated with Hansen solubility theory were identified for each of the polymers used in this study: polystyrene (PS), PMMA, polycarbonate (PC) and PSF. RED, relative energy difference. The blue dots and red crosses denote successful and unsuccessful application of organized stress microfibrillation respectively. The regions within the

dotted lines are magnified in the insets. The numeric labelling of each solvent is detailed in the corresponding table where the weight ratio of components in a mixture is indicated in parentheses, for example, '16(1:5)' refers to a solution of toluene and methanol with a weight ratio of 1:5. THF, tetrahydrofuran.



Extended Data Fig. 5 | The recording of reflected and transmitted colour. **a**, Illustration for experimental setup to record reflected and transmitted colour of polystyrene film (192 kDa) with BDABP printed on a transparent support. **b**, An example showing reflected and transmitted colour of the Kyoto University logo printed on a PET substrate. A camera

(EOS Kiss X5, Canon) with macro-lens (60-mm EFS, Canon) was used to take the photos. The film reflects blue light and transmits the complementary yellow colour, which can be seen in the shadow of the film. The light source is a 190-W ultrahigh-performance mercury lamp (BenQ, China).



Extended Data Fig. 6 | Micro-LED crosslinking of films. **a, b,** The image to be printed is turned into a two-tone image with Floyd–Steinberg dithering using GIMP software (<https://www.gimp.org/>) (**b**). **c,** The image is then converted to a CAD file using Tanner L-Edit IC Layout software (Mentor Graphics, <https://www.mentor.com/tannereda/l-edit>). The micro-LED system can control the on/off state of an array ($1,024 \times 768$) of $1 \mu\text{m} \times 1 \mu\text{m}$ light sources, which produce a single frame. The dashed grids in **c** indicate how such composite images are divided into individual frames. **d,** An SEM image shows the structure and boundary of a pixel

formed in 35-kDa PS/PQ through micro-LED illumination. Scale bar, $1 \mu\text{m}$. **e,** The image resolution is calculated from the printed stripes (light blue) in 35-kDa PS/PQ, which are $1.8 \mu\text{m}$ wide, corresponding to 14,000 dots per inch. The average linewidths were calculated by first converting the image to a binary image, where the light blue and dark blue were mapped to white and black respectively. The average width of the patterned lines is then calculated as the total area of the white pixels divided by the height of the image and total number of lines. Scale bar, $10 \mu\text{m}$.

Turbulent convective length scale in planetary cores

Céline Guervilly^{1*}, Philippe Cardin² & Nathanaël Schaeffer²

Convection is a fundamental physical process in the fluid cores of planets. It is the primary transport mechanism for heat and chemical species and the primary energy source for planetary magnetic fields. Key properties of convection—such as the characteristic flow velocity and length scale—are poorly quantified in planetary cores owing to the strong dependence of these properties on planetary rotation, buoyancy driving and magnetic fields, all of which are difficult to model using realistic conditions. In the absence of strong magnetic fields, the convective flows of the core are expected to be in a regime of rapidly rotating turbulence¹, which remains largely unexplored. Here we use a combination of non-magnetic numerical models designed to explore this regime to show that the convective length scale becomes independent of the viscosity when realistic parameter values are approached and is entirely determined by the flow velocity and the planetary rotation. The velocity decreases very rapidly at smaller scales, so this turbulent convective length scale is a lower limit for the energy-carrying length scales in the flow. Using this approach, we can model realistically the dynamics of small non-magnetic cores such as the Moon. Although modelling the conditions of larger planetary cores remains out of reach, the

fact that the turbulent convective length scale is independent of the viscosity allows a reliable extrapolation to these objects. For the Earth's core conditions, we find that the turbulent convective length scale in the absence of magnetic fields would be about 30 kilometres, which is orders of magnitude larger than the ten-metre viscous length scale. The need to resolve the numerically inaccessible viscous scale could therefore be relaxed in future more realistic geodynamo simulations, at least in weakly magnetized regions.

The very low fluid viscosity in planetary liquid cores implies that the convective flows are turbulent, but this turbulence differs both from three-dimensional (3D) turbulence owing to the anisotropy imposed by the rapid planetary rotation and from two-dimensional (2D) turbulence owing to the presence of Rossby waves². Conditions in planetary cores correspond to small Ekman numbers ($Ek = \nu/\Omega R^2$ with viscosity ν , rotation rate Ω and core radius R), large Reynolds numbers ($Re = UR/\nu$ with flow speed U) and small Rossby numbers ($Ro = U/\Omega R = Re \times Ek$), with, for instance, $Ek \approx 10^{-15}$, $Re \approx 10^9$ and $Ro \approx 10^{-6}$ in the Earth's core³. Numerical models must employ a fluid viscosity that is orders of magnitude larger than realistic values to keep the range of time and length scales involved in the dynamics

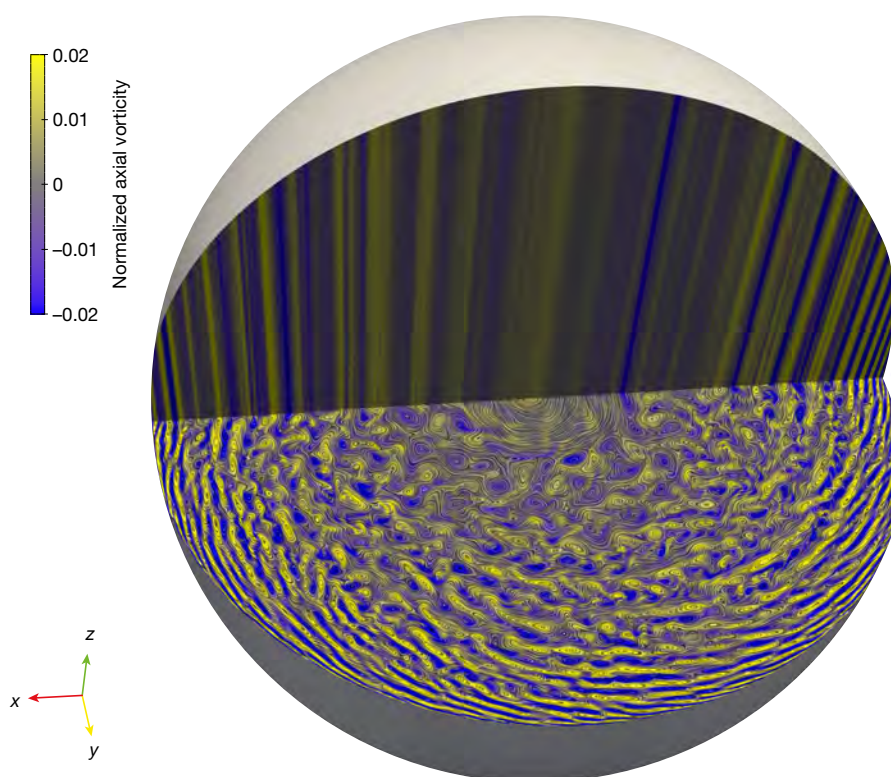


Fig. 1 | Flow in the 3D model. Meridional and equatorial cross-sections of a snapshot of the axial vorticity in the 3D model for $Ek = 10^{-8}$, $Ra = 2 \times 10^{10}$ and $Pr = 10^{-2}$. Streamlines have been superimposed in the equatorial plane. In the colour scale, values of the axial vorticity

are normalized by the planetary vorticity 2Ω . The kinetic energy of the velocity projected on a QG state ($\langle u_s \rangle$, $\langle u_\phi \rangle$, $z\beta\langle u_s \rangle$) in cylindrical polar coordinates (where the angle brackets denote an axial average) is within 0.2% of the total kinetic energy.

¹School of Mathematics, Statistics and Physics, Newcastle University, Newcastle upon Tyne, UK. ²Université Grenoble Alpes, Université Savoie Mont Blanc, CNRS, IRD, IFSTTAR, ISTerre, Grenoble, France. *e-mail: celine.guervilly@ncl.ac.uk

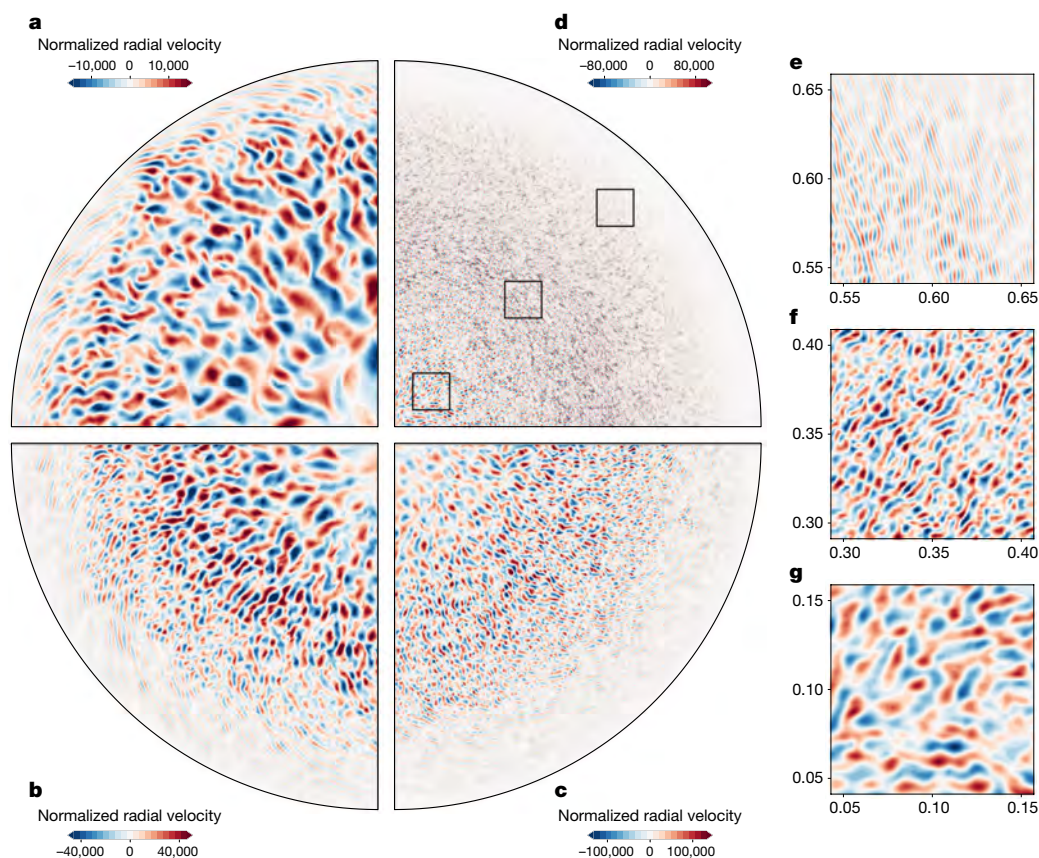


Fig. 2 | Effect of the Rossby number on the flow structure. a–d, Snapshots of the radial velocity in a quarter of the equatorial plane during the statistically steady phase for $Ek = 10^{-8}$ and $Ra = 2.5 \times 10^{10}$ ($Ro = 6 \times 10^{-5}$) (a), $Ek = 10^{-9}$ and $Ra = 2.7 \times 10^{11}$ ($Ro = 2 \times 10^{-5}$) (b), $Ek = 10^{-10}$ and $Ra = 6.3 \times 10^{12}$ ($Ro = 4 \times 10^{-6}$) (c) and $Ek = 10^{-11}$ and $Ra = 5.25 \times 10^{13}$ ($Ro = 3 \times 10^{-7}$) (d). The 3D model was used for a and the QG model

for b–d. e–g, Magnification of the indicated areas for the same parameters as in d. The outer conduction-dominated region, where the dynamics is dominated by Rossby waves (e), and two examples of the inner convective region (f, g). $Pr = 10^{-2}$ in all cases. In the colour scales, the radial velocity is normalized by the viscous velocity scale, ν/R . The axes show the Cartesian coordinates normalized by the sphere radius, R .

manageable, typically⁴ $Ek \geq 10^{-7}$ and $Re \leq 10^4$. This however has the undesirable effect that convection properties are still controlled by the viscosity^{5,6}. In the viscous regime, convection takes the form of tall and narrow columns aligned with the rotation axis with an azimuthal length scale, \mathcal{L}_ν , that depends on the viscosity as $Ek^{1/3}$ (ref. 7), and so $\mathcal{L}_\nu \approx 10$ m for Earth-like parameter values. When nonlinear effects become important in the rapidly rotating turbulent regime of large Re and low Ro , the turbulent convective length scale \mathcal{L}_t is expected to grow above the viscous length scale, up to a scale controlled by the flow velocity^{8–10}. The value of \mathcal{L}_t is currently unknown for planetary cores.

The objective of this work is to provide an estimate of \mathcal{L}_t under core conditions using an extensive numerical exploration of the low-viscosity regime. We use a combination of a state-of-the-art 3D model¹¹ down to $Ek = 10^{-8}$ supplemented by a simplified model of quasi-geostrophic (QG) rotating convection^{12,13} down to $Ek = 10^{-11}$. The simplified QG model takes advantage of the Proudman–Taylor constraint¹⁴ by assuming that the axial vorticity is invariant along the rotation axis. The QG approximation is well supported by the results of the 3D model shown in Fig. 1. The numerical codes solve the governing equations of nonlinear Boussinesq convection driven by homogeneous internal heating in a full sphere geometry (see Methods). Magnetic fields are not included.

For the low Ek values studied here, convection is always in a turbulent state, even near the nonlinear onset^{11,15}, and $Re \geq 10^3$. The convection takes the form of vortical plumes that are radially elongated on scales much shorter than the outer radius (Fig. 2). At large radius, the steepening of the boundary slope inhibits vortical plume convection³. The dynamics there consists mainly of Rossby waves, which appear as elongated vortices with a prograde tilt^{16,17} (Fig. 2e). Their radial

velocities are relatively small so conduction dominates the heat transport in the outer part of the equatorial plane¹⁸. Hereafter, we consider only the dynamics of the inner convective region, which grows wider with increasing Rayleigh number (Ra , which controls the buoyancy driving). The length scale of the convective flows decreases notably with increasing radius (Fig. 2f, g). We find that the convective length scale is controlled by Ro , rather than by any viscous effect. The flows shown in Figs. 1, 2 are snapshots taken once the system has reached a statistically steady state, where the kinetic energy fluctuates around a constant mean value, and are entirely unlike the linear viscous mode at the onset of convection, which consists of drifting columns with narrow azimuthal length scale^{7,19} \mathcal{L}_ν . The convective length scale increases with the buoyancy driving, as seen in the power spectra of the total and radial kinetic energies in Fig. 3. The peak of the radial kinetic energy moves to smaller azimuthal wavenumber m for increasing Ra , as can be observed for the two different Ra shown at $Ek = 10^{-10}$, and is located at a much smaller wavenumber ($m = 133$ and 106 for the smaller and larger Ra , respectively) than the wavenumber of the marginal linear viscous mode at the onset of convection ($m = 258$). Remarkably, the spectra at different Ek and Ra superpose well at wavenumbers larger than the peak, and decrease steeply as m^{-5} (ref. 20). There is therefore a well defined characteristic convective length scale that carries most of the radial kinetic energy; below this scale the velocity amplitude drops very rapidly. This characteristic length scale is thus a limit below which only weak convective motions occur, thereby drastically restricting viscous control and dissipation in the bulk. At wavenumbers smaller than the peak, the velocity becomes anisotropic with a dominant azimuthal component. The kinetic energy is transferred to larger scales, where the dynamics is dominated by

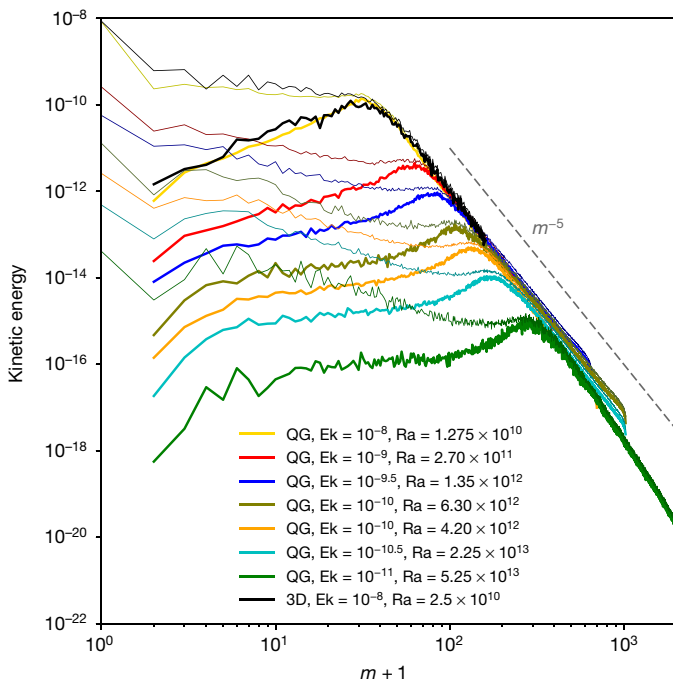


Fig. 3 | Distribution of the kinetic energy at different length scales. Power spectra of the total kinetic energy (thin lines) and radial kinetic energy (thick lines) as a function of the azimuthal wavenumber m at $s = 0.5$ for simulations with different Ek and Ra values for $\text{Pr} = 10^{-2}$ performed with the 3D and QG models. The kinetic energy is averaged in time and normalized by $\rho(\Omega R)^2/2$. The length scale is inversely proportional to m . The dashed line represents a power law with exponent -5 .

propagating Rossby waves, and viscous dissipation occurs in the boundary layers.

In the rapidly rotating turbulent regime, an increase in the convective length scale with the buoyancy driving is expected from scaling arguments^{8–10}, which assume that the production of axial vorticity is governed by a triple inviscid balance between vortex stretching, advection and buoyancy (the so-called Coriolis–inertia–Archimedes balance).

The scaling gives a convective length scale that depends on the flow velocity as $\mathcal{L}_t \propto (\text{Ro}/|\beta|)^{1/2}$, where β is a geometric factor related to the boundary slope (see Methods). This length scale is consistent with the m^{-5} spectra of the kinetic energy. Assuming that the transport in the fluid bulk controls the heat transfer²¹, the scaling uses a balance between the nonlinear advection of temperature and the transport of the mean temperature background to obtain $\text{Re} \propto \text{Ra} \times \text{Ek}/\text{Pr}$, or simply $\text{Ro} \propto \text{Bu}$, where $\text{Bu} = \text{Ra} \times \text{Ek}^2/\text{Pr}$ is the viscosity-free buoyancy parameter. The Prandtl number, Pr , is the ratio of viscosity to thermal diffusivity and is expected to be 0.01 – 0.1 in liquid metal cores. The theoretical scaling law is tested in Fig. 4 against results obtained with the 3D and QG models and against published results obtained with a hybrid model that uses the QG approximation coupled to the 3D temperature¹⁸. The characteristic convective length scale \mathcal{L} corresponds to the peak of the radial kinetic energy spectra. Points obtained at different Ek values collapse onto a single curve, especially for $\text{Ek} < 10^{-9}$, showing that the dependence of the results on the viscosity becomes negligible when core conditions are approached. Importantly, the good agreement obtained between the different numerical models supports the use of the QG approximation for modelling rapidly rotating convection. The data for the velocity and length scale, compensated by their respective theoretical scaling laws, align on a plateau at small Ek values, indicating that the agreement between the simulations and the theoretical scaling improves progressively as Ek decreases. The length scales show little dependence on Pr ; for the velocity scaling law, the exponent is unaffected by Pr but simulations with larger Pr values tend to have a slightly smaller prefactor. To avoid the ‘shingling’ effect that occurs when using diffusion-free parameters²², the scaling of Re is shown in Extended Data Fig. 1 and confirms the overlap of the data for $\text{Ek} \leq 10^{-9}$ and the good agreement with the exponent predicted by the theoretical scaling. The length scale \mathcal{L} corresponds to an azimuthal size in Fig. 4, and we further confirm in Extended Data Fig. 2 that the radial length scale obtained from radial correlations is in good agreement with this azimuthal scale. The radial dependence of the length scale observed in Fig. 2 is also in agreement with the theoretical dependence on $|\beta|^{-1/2}$, as shown in Extended Data Fig. 3. Additional QG simulations performed with differential heating in the presence of an inner core (see Methods) show that the scaling law $\mathcal{L}(\text{Ro})$ of Fig. 4 is valid for other heating modes (Extended Data Fig. 4).

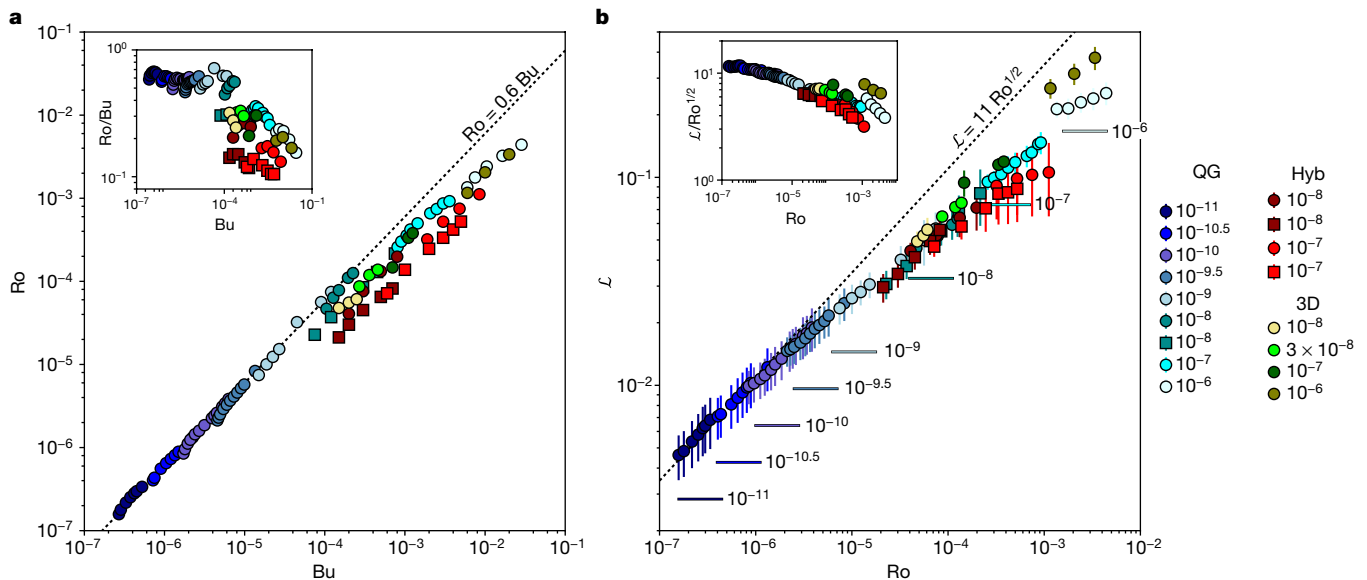


Fig. 4 | Scaling of the velocity and length scale. **a**, Ro as a function of the buoyancy parameter Bu . **b**, Convective length scale \mathcal{L} as a function of Ro using the 3D (green points), QG (blue) and hybrid (red) simulations. Marker colours correspond to Ek (values given in the key) and shapes to Pr (circles, $\text{Pr} = 10^{-2}$ and squares, $\text{Pr} = 10^{-1}$). In **b**, \mathcal{L} is radially averaged

between $s \in [0.1, 0.6]$; the vertical bars give the standard deviation in this interval. The horizontal lines give the linear viscous length scale \mathcal{L}_v at $s = 0.5$ for given Ek and with $\text{Pr} = 10^{-2}$. Insets, the data compensated by the theoretical scaling as a function of Bu (**a**) and Ro (**b**).

The smallest Ekman number computed with our simplified QG numerical model, $Ek = 10^{-11}$, is approximately the value for the core of the Moon²³. Non-magnetic convection in the lunar core is bracketed between the extinction of the dynamo, which occurs at $Re \approx 10^7$, corresponding to a critical magnetic Re of 10 (ref. ²⁴) and the cessation of nonlinear convection, which occurs at $Re \approx 10^3$ (refs ^{11,15}). Between these two events, our results predict that the turbulent length scale decreases as $Ro^{1/2} \propto Ra^{1/2}$ from $0.1R \approx 10$ km to $0.001R \approx 0.1$ km, implying a large reduction in heat transport efficiency (see Methods).

Although non-magnetic, our study has interesting implications for the Earth's core dynamics and geodynamo modelling. Characteristic flow speeds at the core–mantle boundary inferred from the geomagnetic secular variation have $Ro \approx 10^{-6}$ (ref. ²⁵), corresponding to $\mathcal{L}_t \approx 0.01R \approx 30$ km. This value is close to the magnetostrophic cross-over length scale²⁶, a theoretical length scale below which magnetic forces become dynamically important; in the Earth's core, this length scale is estimated to be 1–100 km. The geomagnetic field therefore probably affects core convection. In the presence of magnetic fields, the convective length scale is expected to increase^{27,28}, so the 30-km scale will probably remain a lower limit for the energy-carrying length scales. In the most recent geodynamo simulations^{28–30}, the magnetic field is heterogeneous with a strong dynamical influence in some regions, but not in others. The convection is thus multi-scale, and the dynamics exhibits a viscous dependence ($\mathcal{L}_v \propto Ek^{1/3}$) in the weakly magnetized regions. For more realistic turbulent conditions, we propose that these regions would be in Coriolis–inertia–Archimedes balance instead of viscous balance. This dynamical shift opens up a promising route for more realistic planetary core simulations because the large increase in the characteristic flow length scale from \mathcal{L}_v to \mathcal{L}_t and the steepness of the kinetic energy spectra beyond \mathcal{L}_t permit a relaxation of the numerical resolution constraint in the bulk.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-019-1301-5>.

Received: 28 September 2018; Accepted: 25 April 2019;
Published online 19 June 2019.

- Aurnou, J. et al. Rotating convective turbulence in Earth and planetary cores. *Phys. Earth Planet. Inter.* **246**, 52–71 (2015).
- Vallis, G. K. *Atmospheric and Oceanic Fluid Dynamics: Fundamentals and Large-scale Circulation* (Cambridge Univ. Press, 2006).
- Jones, C. A. in *Treatise on Geophysics 2nd edn* (ed. Schubert, G.) 115–159 (Elsevier, 2015).
- Gastine, T., Wicht, J. & Aubert, J. Scaling regimes in spherical shell rotating convection. *J. Fluid Mech.* **808**, 690–732 (2016).
- King, E. & Buffett, B. Flow speeds and length scales in geodynamo models: the role of viscosity. *Earth Planet. Sci. Lett.* **371–372**, 156–162 (2013).
- Oruba, L. & Dormy, E. Predictive scaling laws for spherical rotating dynamos. *Geophys. J. Int.* **198**, 828–847 (2014).
- Jones, C. A., Soward, A. M. & Mussa, A. I. The onset of thermal convection in a rapidly rotating sphere. *J. Fluid Mech.* **405**, 157–179 (2000).
- Stevenson, D. J. Turbulent thermal convection in the presence of rotation and a magnetic field: a heuristic theory. *Geophys. Astrophys. Fluid Dyn.* **12**, 139–169 (1979).
- Ingersoll, A. P. & Pollard, D. Motion in the interiors and atmospheres of Jupiter and Saturn: scale analysis, anelastic equations, barotropic stability criterion. *Icarus* **52**, 62–80 (1982).
- Aubert, J., Brito, D., Nataf, H.-C., Cardin, P. & Masson, J.-P. A systematic experimental study of rapidly rotating spherical convection in water and liquid gallium. *Phys. Earth Planet. Inter.* **128**, 51–74 (2001).
- Kaplan, E. J., Schaeffer, N., Vidal, J. & Cardin, P. Subcritical thermal convection of liquid metals in a rapidly rotating sphere. *Phys. Rev. Lett.* **119**, 094501 (2017).
- Or, A. C. & Busse, F. H. Convection in a rotating cylindrical annulus. II. Transitions to asymmetric and vacillating flow. *J. Fluid Mech.* **174**, 313–326 (1987).

- Gillet, N., Brito, D., Jault, D. & Nataf, H.-C. Experimental and numerical study of convection in a rapidly rotating spherical shell. *J. Fluid Mech.* **580**, 83–121 (2007).
- Taylor, G. I. The motion of a sphere in a rotating liquid. *Proc. R. Soc. A* **102**, 180–189 (1922).
- Guervilly, C. & Cardin, P. Subcritical convection of liquid metals in a rotating sphere using a quasi-geostrophic model. *J. Fluid Mech.* **808**, 61–89 (2016).
- Miyagoshi, T., Kageyama, A. & Sato, T. Zonal flow formation in the Earth's core. *Nature* **463**, 793–796 (2010).
- Sumita, I. & Olson, P. Experiments on highly supercritical thermal convection in a rapidly rotating hemispherical shell. *J. Fluid Mech.* **492**, 271–287 (2003).
- Guervilly, C. & Cardin, P. Multiple zonal jets and convective heat transport barriers in a quasi-geostrophic model of planetary cores. *Geophys. J. Int.* **211**, 455–471 (2017).
- Zhang, K. Spiralling columnar convection in rapidly rotating spherical fluid shells. *J. Fluid Mech.* **236**, 535–556 (1992).
- Schaeffer, N. & Cardin, P. Rossby-wave turbulence in a rapidly rotating sphere. *Nonlinear Process. Geophys.* **12**, 947–953 (2005).
- Julien, K., Knobloch, E., Rubio, A. & Vasil, G. Heat transport in low-Rossby-number Rayleigh–Bénard convection. *Phys. Rev. Lett.* **109**, 254503 (2012).
- Cheng, J. S. & Aurnou, J. M. Tests of diffusion-free scaling behaviors in numerical dynamo datasets. *Earth Planet. Sci. Lett.* **436**, 121–129 (2016).
- Weber, R. C., Lin, P.-Y., Garnero, E. J., Williams, Q. & Lognonne, P. Seismic detection of the lunar core. *Science* **331**, 309–312 (2011).
- Christensen, U. R. & Aubert, J. Scaling properties of convection-driven dynamos in rotating spherical shells and application to planetary magnetic fields. *Geophys. J. Int.* **166**, 97–114 (2006).
- Holme, R. & Olsen, N. Core surface flow modelling from high-resolution secular variation. *Geophys. J. Int.* **166**, 518–528 (2006).
- Aurnou, J. & King, E. The cross-over to magnetostrophic convection in planetary dynamo systems. *Proc. R. Soc. A* **473**, 20160731 (2017).
- Chandrasekhar, S. *Hydrodynamic and Hydromagnetic Stability* (Clarendon, 1961).
- Yadav, R., Gastine, T., Christensen, U., Wolk, S. J. & Poppenhaeger, K. Approaching a realistic force balance in geodynamo simulations. *Proc. Natl Acad. Sci. USA* **113**, 12065–12070 (2016).
- Aubert, J., Gastine, T. & Fournier, A. Spherical convective dynamos in the rapidly rotating asymptotic regime. *J. Fluid Mech.* **813**, 558–593 (2017).
- Schaeffer, N., Jault, D., Nataf, H.-C. & Fournier, A. Turbulent geodynamo simulations: a leap towards Earth's core. *Geophys. J. Int.* **211**, 1–29 (2017).

Acknowledgements C.G. was supported by the UK Natural Environment Research Council under grant NE/M017893/1. P.C. and N.S. were supported by the French Agence Nationale de la Recherche under grants ANR-13-BS06-0010 (TuDy) and ANR-14-CE33-0012 (MagLune). N.S. acknowledges GENCI for access to the Occigen resource (CINES) under grants A0020407382 and A0040407382. This research made use of the Rocket High Performance Computing service at Newcastle University, the ARCHER UK National Supercomputing Service (<http://www.archer.ac.uk>), and the DiRAC@Durham facility managed by the Institute for Computational Cosmology on behalf of the STFC DiRAC HPC Facility (<http://www.dirac.ac.uk>) and funded by BEIS capital funding via STFC capital grants ST/P002293/1, ST/R002371/1 and ST/S002502/1, and Durham University and STFC operations grant ST/R000832/1. Some computations were also performed on the Froggy platform of CIMENT (<https://ciment.ujf-grenoble.fr>), supported by the Rhône-Alpes region (CPER07_13 CIRA), OSUG@2020 LabEx (ANR10 LABX56) and Equip@Meso (ANR10 EQPX-29-01). ISTERre is part of Labex OSUG@2020 (ANR10 LABX56).

Reviewer information Nature thanks Bruce Buffett and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions C.G. and P.C. performed the numerical simulations with the QG code. N.S. performed the numerical simulations with the 3D code. All authors contributed to the analysis of the data and the preparation of the manuscript.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-019-1301-5>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-019-1301-5>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to C.G.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

METHODS

We model Boussinesq convection driven by homogeneous internal heating in a full-sphere geometry. This problem is relevant for planetary cores without a solid inner core, and is thus relevant for most of the Earth's history³¹. The model does not include magnetic fields. The sphere rotates at a rate Ω around the axis directed along the unit vector \mathbf{e}_z . The acceleration due to gravity \mathbf{g} is radial and increases linearly with the radius r as $\mathbf{g} = -g_0 r \mathbf{e}_r$. The governing equations are written in a dimensionless form and are obtained by scaling length by R , scaling time by R^2/ν and scaling temperature by $\nu SR^2/(6\rho C_p \kappa^2)$, where R is the outer radius of the core, ν the fluid kinematic viscosity, S the internal volumetric heating, κ the thermal diffusivity, ρ the density and C_p is the heat capacity at constant pressure. The dimensionless numbers are the Ekman number, $\text{Ek} = \nu/(\Omega R^2)$, the Rayleigh number, $\text{Ra} = \alpha g_0 SR^6/(6\rho C_p \nu \kappa^2)$, where α is the thermal expansion coefficient, and the Prandtl number, $\text{Pr} = \nu/\kappa$. This study focuses on Pr values smaller than unity, which are relevant for the thermal convection of liquid metal cores³². The system of dimensionless equations is:

$$\frac{\partial \mathbf{u}}{\partial t} + (\mathbf{u} \cdot \nabla) \mathbf{u} + \frac{2}{\text{Ek}} \mathbf{e}_z \times \mathbf{u} = -\nabla p + \nabla^2 \mathbf{u} + \text{Ra} \Theta \mathbf{r} \quad (1)$$

$$\nabla \cdot \mathbf{u} = 0 \quad (2)$$

$$\frac{\partial \Theta}{\partial t} + \mathbf{u} \cdot \nabla \Theta - \frac{2}{\text{Pr}} r u_r = \frac{1}{\text{Pr}} \nabla^2 \Theta \quad (3)$$

where \mathbf{u} is the velocity field, p the pressure and Θ the temperature perturbation relative to the static temperature $T_{\text{static}} = (1-r^2)/\text{Pr}$. We use no-slip boundary conditions and a fixed temperature at the outer boundary.

3D numerical model. For the 3D simulations, we use the code XSHELLS¹¹, which solves equations (1)–(3) using finite differences in the radial direction and spherical harmonic expansion³³. The input parameters and numerical resolutions used for the 3D simulations are given in the Supplementary Information. In the 3D simulations, Pr is fixed at 10^{-2} and Ek is varied between 10^{-6} and 10^{-8} . The most computationally demanding simulations performed at $\text{Ek} = 10^{-8}$ were run with a numerical resolution of 2,016 radial grid points and truncation degree $L = 351$, and order $M = 319$ for the spherical harmonics. Hyperviscosity was used in all the 3D simulations, with viscosity depending on spherical harmonic degree, ℓ , only when ℓ is greater than a cut-off value $\ell_c = 0.9L$ (ref. ¹¹). For $\ell < \ell_c$, $\nu(\ell) = \nu_0$, and for $\ell \geq \ell_c$, $\nu(\ell) = \nu_0 q^{\ell - \ell_c}$. We set $q = (\nu_{\text{max}}/\nu_0)^{1/(L - \ell_c)}$ and $\nu_{\text{max}} \leq 100$.

QG numerical model. For simulations at smaller Ek values, we assume that the rotational constraint is such that the variations of the velocity along the axial direction are small compared with the variations along the orthogonal directions. We use a QG approximation for rapidly rotating spherical convection developed from the Busse³⁴ annulus model^{12,35} and widely used in the context of planetary core convection^{36–40}. The dynamics are assumed to be dominated by the geostrophic balance, that is, the Coriolis force balances the pressure gradient at leading order. The leading-order velocity \mathbf{u}_\perp is invariant along z and $\mathbf{u}_\perp = (u_s, u_\phi, 0)$ in cylindrical polar coordinates. QG convection is driven by the cylindrical component of gravity $-g_0 s$. By taking the axial average of the z component of the curl of the Navier–Stokes equation, we obtain the equation for the leading-order axial vorticity, ζ :

$$\frac{\partial \zeta}{\partial t} + (\mathbf{u}_\perp \cdot \nabla_\perp) \zeta - \left(\frac{2}{\text{Ek}} + \zeta \right) \left\langle \frac{\partial u_z}{\partial z} \right\rangle = \nabla_\perp^2 \zeta - \text{Ra} \left\langle \frac{\partial \Theta}{\partial \phi} \right\rangle \quad (4)$$

where $\nabla_\perp f \equiv (\partial_s f, \partial_\phi f, 0)$, $\nabla_\perp^2 f \equiv \partial_s^2 f + s^{-1} \partial_s f + s^{-2} \partial_\phi^2 f$, and the angle brackets denote an axial average between $-H$ and $+H$, where $H = \sqrt{1-s^2}$ is the height of the spherical boundary from the equatorial plane.

The velocity is described by a streamfunction ψ that models the non-axisymmetric (that is, ϕ -dependent) components with the addition of an axisymmetric azimuthal flow, \bar{u}_ϕ , where the overbar denotes an azimuthal average:

$$\mathbf{u}_\perp = \frac{1}{H} \nabla \times (H \psi \mathbf{e}_z) + \bar{u}_\phi \mathbf{e}_\phi \quad (5)$$

This choice of the streamfunction accounts for mass conservation at the outer boundary⁴¹. We assume that the axial velocity u_z is linear in z and has two contributions: the main contribution comes from mass conservation at the outer boundary and is proportional to $\beta = H'/H$; the other contribution accounts for Ekman pumping, which is produced by the viscous boundary layer and scales as $\text{Ek}^{1/2}$. The Ekman pumping is parameterized by the formula obtained by asymptotic methods in the limit of small Ek for a linear Ekman layer⁴².

The streamfunction ψ describes only the non-axisymmetric motions, so \bar{u}_ϕ is obtained by taking the azimuthal and axial averages of the ϕ component of the Navier–Stokes equation to give:

$$\frac{\partial \bar{u}_\phi}{\partial t} + \bar{u}_s \frac{\partial \bar{u}_\phi}{\partial s} + \frac{\bar{u}_s \bar{u}_\phi}{s} = \nabla_\perp^2 \bar{u}_\phi - \frac{\bar{u}_\phi}{s^2} - \frac{1}{\text{Ek}^{1/2} H^{3/2}} \bar{u}_\phi \quad (6)$$

where the last term on the right-hand side corresponds to the Coriolis term simplified using mass conservation³⁶.

The equation for the temperature perturbation Θ in the QG model is obtained by taking the axial average of the temperature equation and assuming that Θ is invariant along z to obtain:

$$\frac{\partial \Theta}{\partial t} + \mathbf{u} \cdot \nabla_\perp \Theta - \frac{4}{3 \text{Pr}} s u_s = \frac{1}{\text{Pr}} \nabla_\perp^2 \Theta \quad (7)$$

We use the gradient of the z -averaged static temperature profile, $(T_{\text{static}})' = -4s/3\text{Pr}$, rather than the gradient of the z -invariant static temperature profile, $(T_{\text{static}}^{2D})' = -3s/\text{Pr}$, to allow for a direct comparison of the Rayleigh numbers used in the different models. The assumption that Θ is invariant along z is not rigorously justified and is used for numerical convenience; it permits us to treat the numerical problem in two dimensions, considerably reducing the computational load. The evolution equation for the streamfunction, the axisymmetric velocity and the temperature are solved on a 2D grid in the equatorial plane. The QG code uses a pseudo-spectral code with a Fourier decomposition in the azimuthal direction and a second-order finite-difference scheme in the radius, with irregular spacing. The input parameters and the numerical resolutions used for the QG simulations are given in the Supplementary Information. Pr is varied between 10^{-1} and 10^{-2} and Ek is varied between 10^{-6} and 10^{-11} , allowing an overlap with the 3D simulations over two decades in Ek . The most computationally demanding simulations performed at $\text{Ek} = 10^{-11}$ were run with a numerical resolution of 4,000 radial grid points and 2,048 Fourier modes in the azimuth.

The influence of the assumption of the z invariance of Θ on the QG results is tested by comparing the QG results with published results obtained with a hybrid QG–3D model¹⁸ for $\text{Ek} \in [10^{-8}, 10^{-7}]$. In the hybrid model, the temperature is solved in 3D and coupled to the QG implementation for the velocity. Figure 4 shows good agreement obtained between the QG and hybrid results for overlapping parameters, demonstrating that although this assumption is not mathematically justified, it does not substantially influence QG convection.

QG model with differential heating. To test the dependence of our results on the heating mode and the presence of an inner core, we performed additional QG simulations using differential heating with fixed temperature boundary conditions and an inner core of radius $R_i = 0.35$. The temperature is scaled by $\text{Pr} \Delta T$. The equation for the temperature perturbation is solved in two dimensions with a static temperature gradient $(T_{\text{static}}^{2D})' = \gamma / (\text{Pr} \times \ln(R_i/s))$, where the constant $\gamma = 0.445$ is used to re-scale the z -invariant temperature profile so that it corresponds closely with the z -averaged static temperature profile³⁹. This re-scaling allows us to compare directly the Ra used in the QG simulations with those used in the 3D models. Pr is varied between 1 and 10^{-2} and Ek is varied between 10^{-8} and 10^{-10} . The input parameters and the numerical resolutions used for the QG simulations with differential heating are given in the Supplementary Information. Extended Data Fig. 4 shows that, for these additional simulations, the azimuthal convective length scale as a function of Ro follows closely the scaling law derived for the QG simulations with internal heating and without an inner core.

Definition of the output parameters. The simulations are started from either a small temperature perturbation or the snapshot of a previous simulation performed at a different Rayleigh number in order to minimize the transient phase before saturation. All simulations are run to saturation, as shown in Extended Data Fig. 5, where we plot the time series of the kinetic energy density for one representative QG case at $\text{Ek} = 10^{-11}$ and one representative 3D case at $\text{Ek} = 10^{-8}$. For consistency, the kinetic energy density K is defined in both cases as:

$$K = \frac{1}{2V} \int (u_s^2 + u_\phi^2) dV \quad (8)$$

where V is the volume of the sphere, and the kinetic energy density of the axisymmetric velocity is:

$$K_{\text{axi}} = \frac{1}{2V} \int (\bar{u}_\phi)^2 dV \quad (9)$$

A number of output parameters are given in the Supplementary Information. The characteristic velocity \mathcal{U} used to calculate Ro and Re is based on the root mean square of the radial velocity averaged in volume and time over at least ten convective turnover timescales.

The convective length scale is calculated as $\mathcal{L}(s) = \pi s / m_p(s)$, where m_p is the wavenumber at the peak of the radial kinetic energy spectrum. The peak is determined by smoothing the time-averaged radial kinetic energy spectra with a polynomial of degree 14.

The radial length scale of the convective flow $\mathcal{L}_r(s)$ is calculated using the auto-correlation function f of the radial component of the velocity field. For a given radius s , we calculate:

$$f(\text{ds}) = \overline{u_s(s, \phi, t) u_s(s + \text{ds}, \phi, t)} \quad (10)$$

where the overbar denotes an azimuthal average. Snapshots covering at least two dynamical timescales are used to compute the temporal average. $\mathcal{L}_r(s)$ is the full-width at half-maximum of f .

Inviscid scaling theory. The theoretical scaling of the velocity and length scale^{3,4,8–10} assumes a triple inviscid balance in the axial vorticity equation between the vorticity advection, vortex stretching and vorticity generation by buoyancy:

$$\frac{\text{Re}^2}{\mathcal{L}_t^2} \approx \frac{|\beta|\text{Re}}{\text{Ek}} \approx \frac{\text{Ra}\mathcal{T}}{\mathcal{L}_t} \quad (11)$$

where \mathcal{T} denotes a typical temperature perturbation and we assume that the typical axial vorticity is Re/\mathcal{L}_t . The turbulent convective length scale then scales as $\mathcal{L}_t \propto (\text{Ro}/|\beta|)^{1/2}$. Assuming that, in rapidly rotating convection, the heat transfer is controlled by the transport in the bulk of the fluid rather than in the thermal boundary layers²¹, we obtain a balance between the nonlinear advection of heat and the transport of the mean temperature background in the temperature equation³:

$$\frac{\text{Re}\mathcal{T}}{\mathcal{L}_t} \approx \frac{s\text{Re}}{\text{Pr}} \quad (12)$$

Combining equations (11) and (12) leads to $\text{Re} \propto \text{Ra} \times \text{Ek}/\text{Pr}$, where we neglect the geometric term $s/|\beta|$. The efficiency of the heat transport can be measured by the ratio q/q_s , where both the convective heat flux $q = \text{Pr}\mathcal{T}\text{Re}$ and the static heat flux $q_s \propto 1/\text{Pr}$ are dimensionless. The theoretical scalings of the velocity and temperature perturbation imply that $q/q_s \propto \mathcal{L}_t^3 \times \text{Pr}/\text{Ek}$.

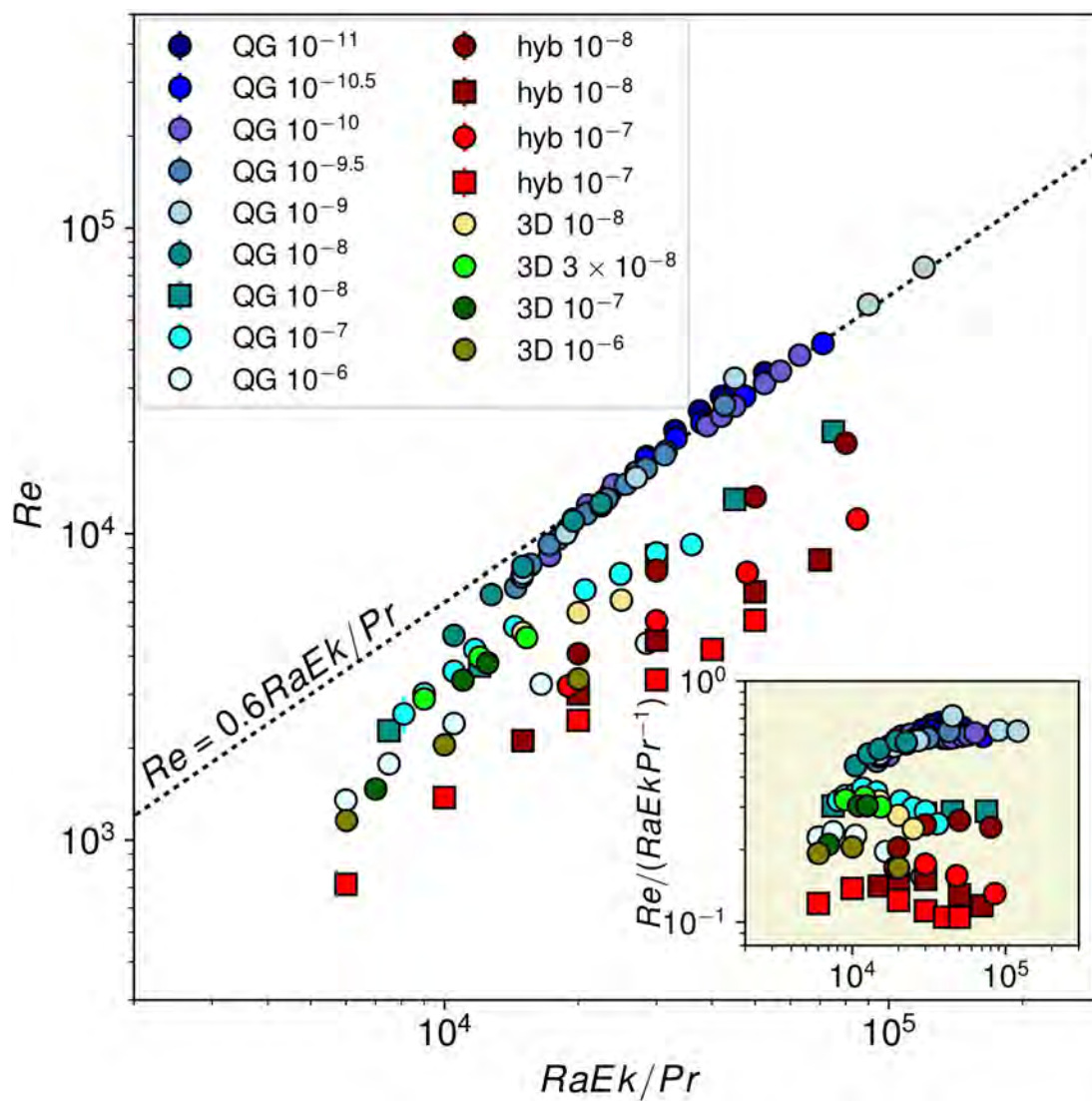
Data availability

Source data for Figs. 3, 4 are provided with this paper. The data generated during this study are included in the Supplementary Information file. Any additional data that support the findings of this study are available from the corresponding author on reasonable request.

Code availability

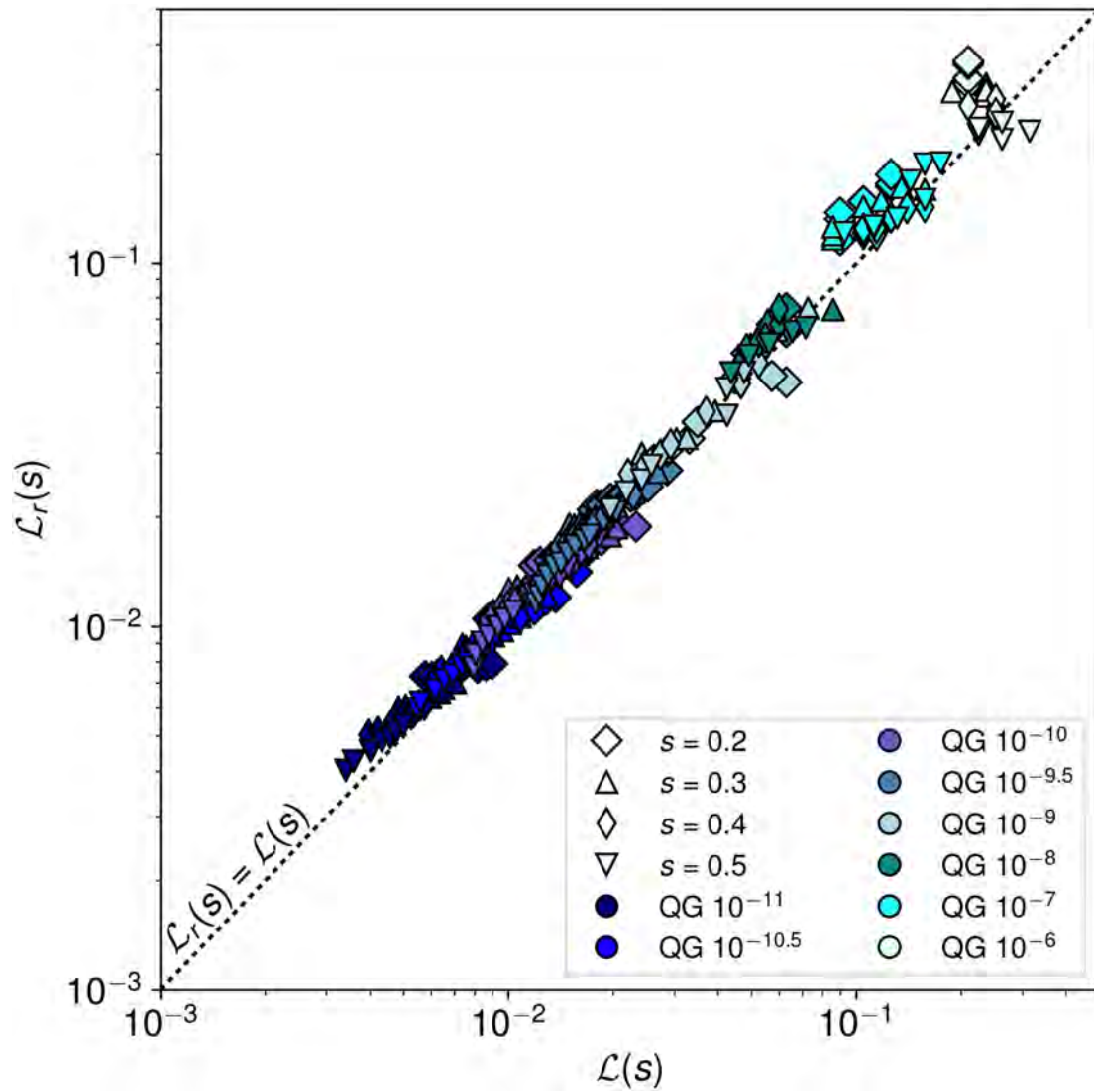
The 3D numerical code XSHELLS is freely available at <https://bitbucket.org/nschaeff/xshells> and is distributed under the open source CeCILL License (http://www.cecill.info/licences/Licence_CeCILL_V2.1-en.html). The QG numerical code is available from the corresponding author on request.

31. Labrosse, S. Thermal evolution of the core with a high thermal conductivity. *Phys. Earth Planet. Inter.* **247**, 36–55 (2015).
32. Pozzo, M., Davies, C., Gubbins, D. & Alfe, D. Thermal and electrical conductivity of iron at Earth's core conditions. *Nature* **485**, 355–358 (2012).
33. Schaeffer, N. Efficient spherical harmonic transforms aimed at pseudospectral numerical simulations. *Geochem. Geophys. Geosyst.* **14**, 751–758 (2013).
34. Busse, F. H. Thermal instabilities in rapidly rotating systems. *J. Fluid Mech.* **44**, 441–460 (1970).
35. Cardin, P. & Olson, P. Chaotic thermal convection in a rapidly rotating spherical shell: consequences for flow in the outer core. *Phys. Earth Planet. Inter.* **82**, 235–259 (1994).
36. Aubert, J., Gillet, N. & Cardin, P. Quasigeostrophic models of convection in rotating spherical shells. *Geochem. Geophys. Geosyst.* **4**, 1052 (2003).
37. Morin, V. & Dormy, E. Time dependent beta-convection in rapidly rotating spherical shells. *Phys. Fluids* **16**, 1603–1609 (2004).
38. Plaut, E., Lebranchu, Y., Simitev, R. & Busse, F. H. On the Reynolds stresses and mean fields generated by pure waves: applications to shear flows and convection in a rotating shell. *J. Fluid Mech.* **602**, 303–326 (2008).
39. Gillet, N. & Jones, C. A. The quasi-geostrophic model for rapidly rotating spherical convection outside the tangent cylinder. *J. Fluid Mech.* **554**, 343–369 (2006).
40. Calkins, M., Aurnou, J., Eldredge, J. & Julien, K. The influence of fluid properties on the morphology of core turbulence and the geomagnetic field. *Earth Planet. Sci. Lett.* **359–360**, 55–60 (2012).
41. Schaeffer, N. & Cardin, P. Quasigeostrophic model of the instabilities of the Stewartson layer in flat and depth-varying containers. *Phys. Fluids* **17**, 104111 (2005).
42. Greenspan, H. P. *The Theory of Rotating Fluids* (Cambridge Univ. Press, 1968).



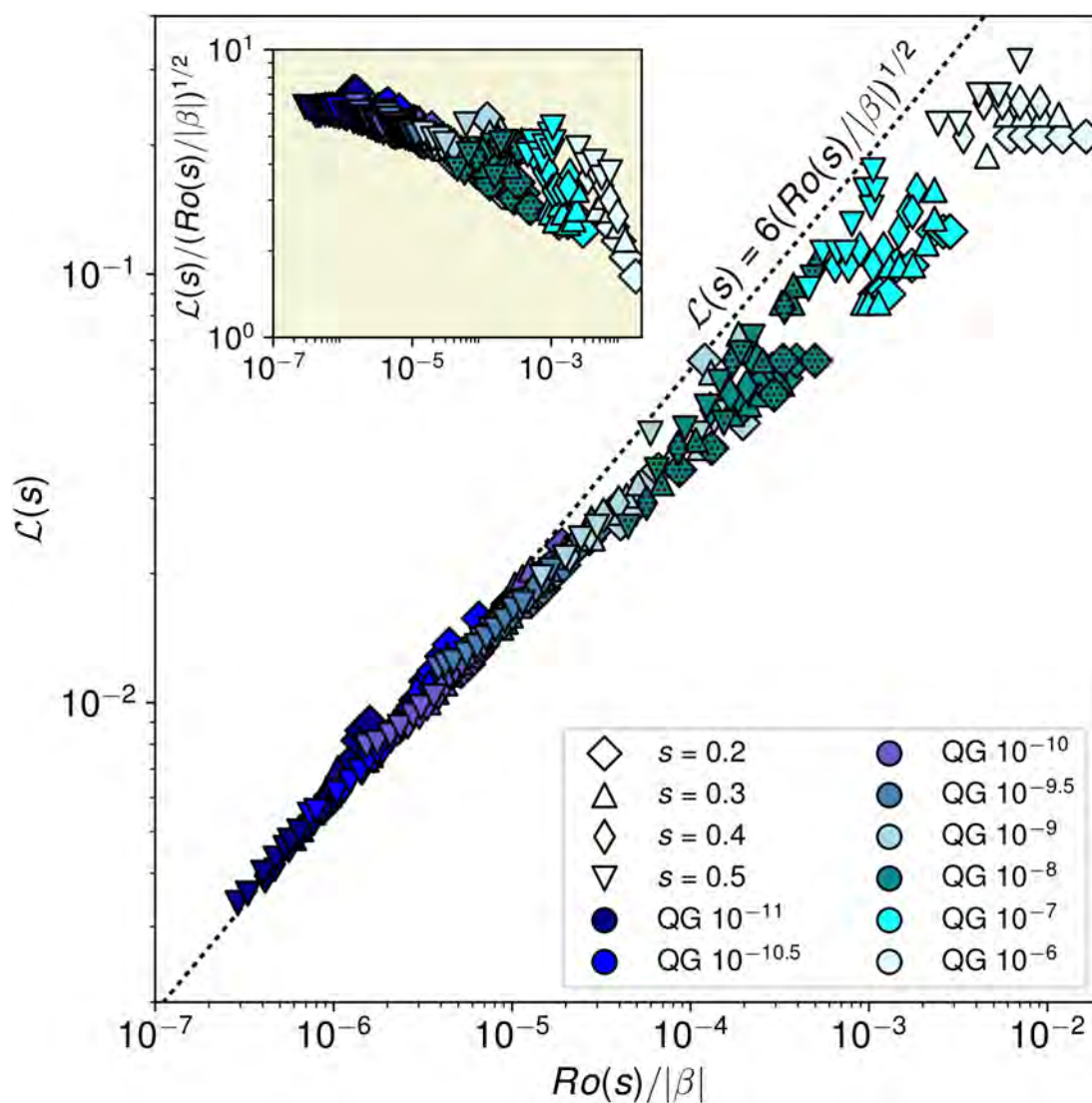
Extended Data Fig. 1 | Scaling of the Reynolds number. Re as a function of $Ra \times Ek/Pr$ in simulations performed with the 3D model (green data points) for $Ek \in [10^{-8}, 10^{-6}]$, the QG model (blue data points) for $Ek \in [10^{-11}, 10^{-6}]$, and the hybrid model (red data points) for $Ek \in [10^{-8}, 10^{-7}]$. Marker colours correspond to Ekman numbers (values given in the

key) and marker shapes correspond to Prandtl numbers (circles, $Pr = 10^{-2}$ and squares, $Pr = 10^{-1}$). The dashed line represents $Re = 0.6Ra \times Ek/Pr$. Inset, the same data compensated by theoretical scaling as a function of $Ra \times Ek/Pr$.



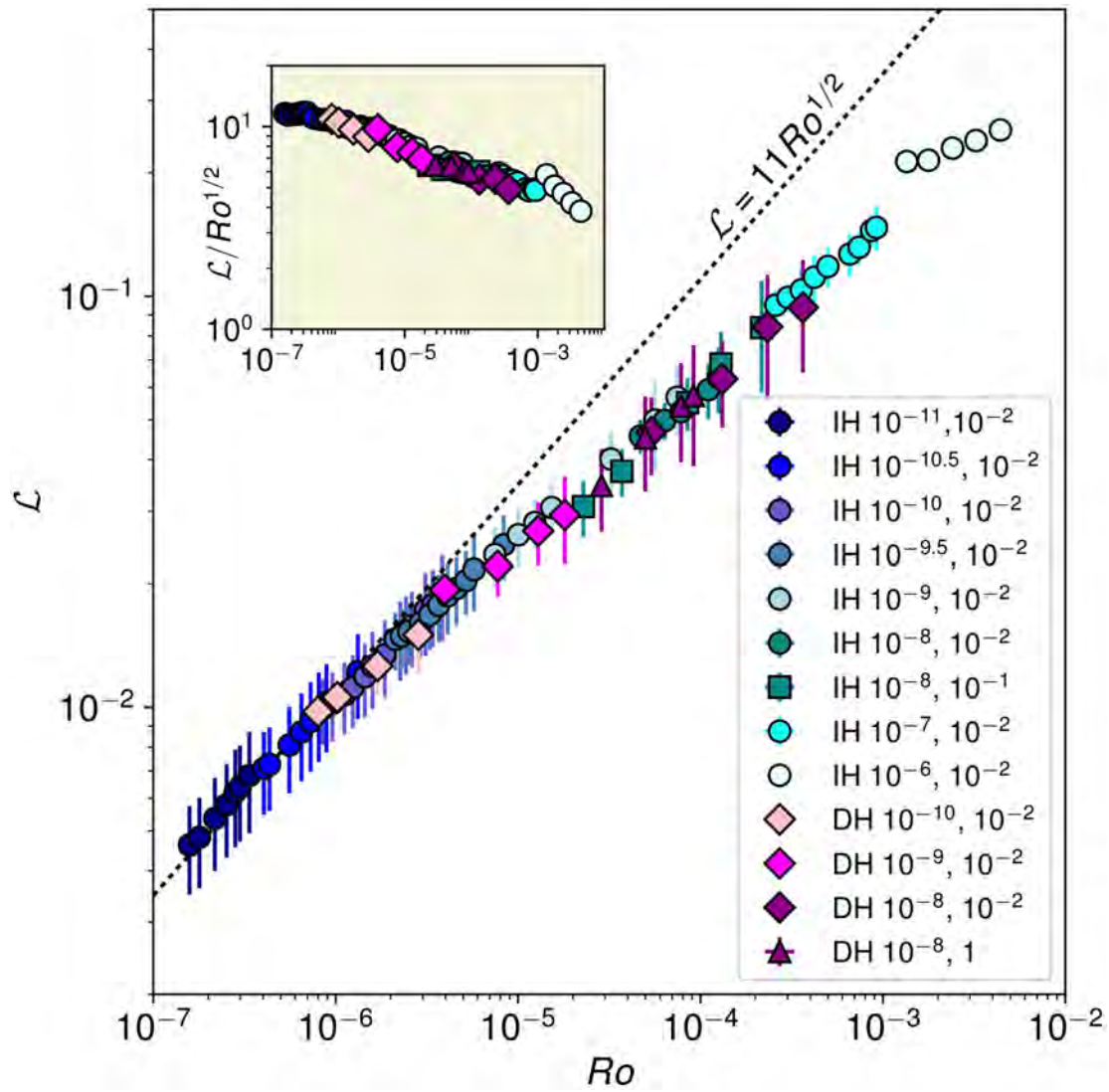
Extended Data Fig. 2 | Comparison of the radial length scale with the azimuthal length scale. Radial scale of the convective flows $\mathcal{L}_r(s)$ as a function of the azimuthal length scale $\mathcal{L}(s)$ obtained with the QG model at different radii s . Marker colours correspond to Ekman numbers (with $\text{Pr} = 10^{-2}$) and marker shapes correspond to the given radii. The radial

scale is calculated from auto-correlation functions of the radial velocity, and the convective length scale corresponds to an azimuthal scale calculated from the peak of the power spectra of the radial kinetic energy at radius s . The dashed line represents $\mathcal{L}_r(s) = \mathcal{L}(s)$.



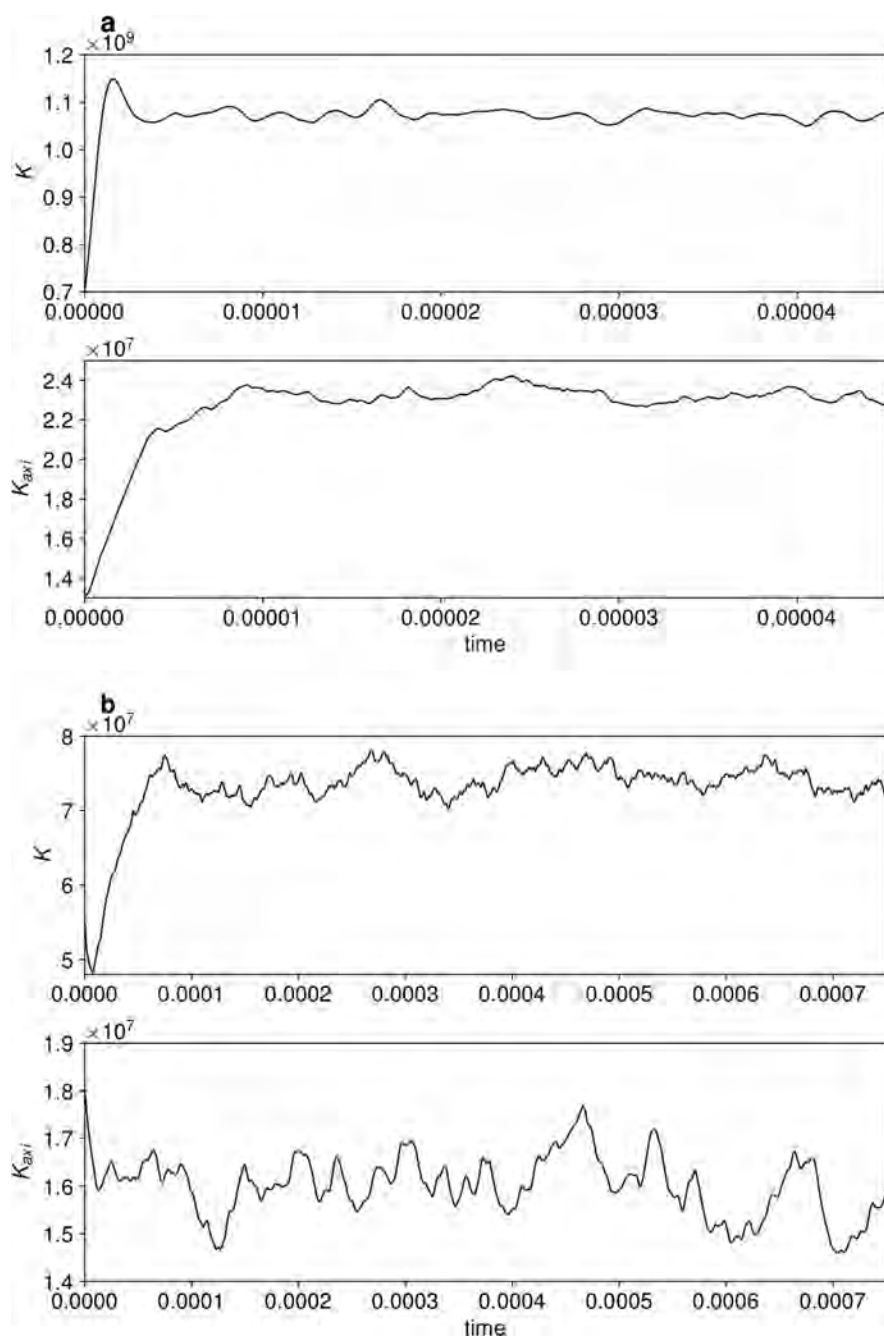
Extended Data Fig. 3 | Variation of the convective length scale with radius. Convective length scale $\mathcal{L}(s)$ as a function of $Ro(s)/|\beta|$ obtained with the QG model at different radii s . Marker colours correspond to Ekman numbers, solid-colour markers correspond to $Pr = 0.01$, dotted markers to $Pr = 0.1$, and marker shapes correspond to the given radii.

The convective length scale corresponds to an azimuthal scale calculated from the peak of the power spectra of the radial kinetic energy at radius s . The dashed line represents $\mathcal{L}_r(s) = 6(Ro(s)/|\beta|)^{1/2}$. Inset, the length scale compensated by theoretical scaling as a function of $Ro(s)/|\beta|$.



Extended Data Fig. 4 | Effect of the heating mode on the convective length scale. Convective length scale \mathcal{L} as a function of Ro obtained with the QG model for internal heating (IH, same points as in Fig. 4) and differential heating (DH) with an inner core of radius $R_i = 0.35$. $Ek \in [10^{-11}, 10^{-6}]$ and $Pr \in \{10^{-2}, 10^{-1}, 1\}$ are given in the key.

The convective scale is averaged over radii between $s = 0.1$ and 0.6 and the vertical error bars give the standard deviation in this interval. The dashed line represents $\mathcal{L} = 11Ro^{1/2}$. Inset, the same data compensated by theoretical scaling as a function of Ro .



Extended Data Fig. 5 | Time series of the kinetic energy density for two representative simulations. a, b, Time series of the kinetic energy density K and the kinetic energy density of the axisymmetric flow K_{axi} for

$Ek = 10^{-11}$, $Pr = 0.01$ and $Ra = 3.75 \times 10^{13}$ using the QG model (a) and $Ek = 10^{-8}$, $Pr = 0.01$ and $Ra = 2 \times 10^{10}$ using the 3D model (b). Time is given in units of a viscous timescale.

Global change drives modern plankton communities away from the pre-industrial state

Lukas Jonkers^{1*}, Helmut Hillebrand^{2,3,4} & Michal Kucera¹

The ocean—the Earth’s largest ecosystem—is increasingly affected by anthropogenic climate change^{1,2}. Large and globally consistent shifts have been detected in species phenology, range extension and community composition in marine ecosystems^{3–5}. However, despite evidence for ongoing change, it remains unknown whether marine ecosystems have entered an Anthropocene⁶ state beyond the natural decadal to centennial variability. This is because most observational time series lack a long-term baseline, and the few time series that extend back into the pre-industrial era have limited spatial coverage^{7,8}. Here we use the unique potential of the sedimentary record of planktonic foraminifera—ubiquitous marine zooplankton—to provide a global pre-industrial baseline for the composition of modern species communities. We use a global compilation of 3,774 seafloor-derived planktonic foraminifera communities of pre-industrial age⁹ and compare these with communities from sediment-trap time series that have sampled plankton flux since ad 1978 (33 sites, 87 observation years). We find that the Anthropocene assemblages differ from their pre-industrial counterparts in proportion to the historical change in temperature. We observe community changes towards warmer or cooler compositions that are consistent with historical changes in temperature in 85% of the cases. These observations not only confirm the existing evidence for changes in marine zooplankton communities in historical times, but also demonstrate that Anthropocene communities of a globally distributed zooplankton group systematically differ from their unperturbed pre-industrial state.

To determine whether anthropogenic climate change has affected the marine environment beyond its natural state, it is essential to compare modern observations to a pre-industrial baseline. Because such a baseline is available for the physical state of the ocean, it has been established that, in response to global warming, the sea-surface temperature field has changed significantly since the onset of industrialization approximately 170 years ago^{1,2} (Fig. 1). Marine ecosystem research, on the other hand, is almost exclusively based on observations since the mid-twentieth century and the pre-industrial baseline is therefore mostly unknown. Although existing observations provide strong evidence for changes in marine ecosystems in a direction that is consistent with late-twentieth century climate change^{3,4,10,11}, this lack of a pre-industrial reference prevents assessing the degree to which Anthropocene marine ecosystems differ from their natural, pre-industrial state¹². This affects our ability to predict the effects of global change on marine ecosystem functioning and the resulting impacts on the resources that they provide to society.

Planktonic foraminifera are a globally ubiquitous group of marine zooplankton. Their distribution is primarily controlled by temperature^{13,14}. About 40 morphospecies are known¹⁵; they occur most abundantly in the surface mixed layer, but some species can be found alive down to several hundreds of metres¹⁶. They are unique among marine zooplankton because their calcite shells are well-preserved in marine sediments. This renders them an ideal model system to investigate the

influence of global change on marine zooplankton, because seafloor sediments offer the chance to obtain an accurate picture of the composition of planktonic foraminifera communities in the past. Indeed, their skeletal remains have extensively been used to elucidate past climate and ecological changes^{17,18}. However, the influence of anthropogenic climate change on planktonic foraminifera communities has only been investigated in very few studies with a regional focus^{7,19} and an assessment on a global scale is lacking.

Here we use a quality-controlled compilation of planktonic foraminifera species assemblages of pre-industrial age from seafloor sediments (see Methods) and compare these to modern (1978–2013) assemblages based on observations from moored sediment-trap time series (see Methods). The sediment samples ($n = 3,754$) cover all major ecological provinces, and almost the entire global temperature gradient, at high resolution (Fig. 1). By virtue of the sedimentation process and sediment mixing by deep-sea organisms, foraminiferal shells that are extracted from the uppermost sediment layer represent a centennially to millennially integrated assemblage before marked human influences (see Methods). The modern assemblages are based on the integration of shell flux of planktonic foraminifera over at least one year and are thus not affected by ontogeny and/or seasonality. Most of the sediment-trap sites are from the Northern Hemisphere; however, they cover the global thermal gradient and include time series spanning up to 12 years. We compare these Anthropocene assemblages with those in the sediment using a square-chord distance metric on species relative abundances (see Methods). We find that all Anthropocene communities differ from those in the nearest sediment sample and that the degree of dissimilarity scales with the temperature change since ad 1870 at each site ($r = 0.53$, $P = 0.001$, $n = 33$; Fig. 2a). This suggests that planktonic foraminifera communities have changed considerably since the pre-industrial period, and that they have done so in proportion to the magnitude of local temperature change. When comparing the modern species composition with the sediment samples, we find that for each modern assemblage the most similar sedimentary analogue is not the assemblage from the nearest core top, but an assemblage from a core top located elsewhere (Fig. 2b). Thus, the changes in community composition and presence of close analogues elsewhere indicate a directional shift at the community level, rather than a random reshuffling of the species forming previously unseen communities. On the basis of the difference between the modern and pre-industrial communities, we estimate that this shift equates to a median latitudinal displacement of 602 km (range, 45–2,557 km) since pre-industrial times (Fig. 2c).

To evaluate the direction of change in community composition, we consider the location of sedimentary assemblages with species compositions that are most similar to the Anthropocene assemblages from the sediment traps. We observe that these sediment assemblages are in most cases from warmer areas, thus confirming that the twentieth-century community composition shows an imprint of global warming (Fig. 3). Warming signatures are found across the globe and in a range

¹MARUM - Center for Marine Environmental Sciences, University of Bremen, Bremen, Germany. ²Institute for Chemistry and Biology of Marine Environments, Carl von Ossietzky University, Oldenburg, Wilhelmshaven, Germany. ³Helmholtz Institute for Functional Marine Biodiversity at the University of Oldenburg, Oldenburg, Germany. ⁴Alfred Wegener Institute, Helmholtz-Centre for Polar and Marine Research, Bremerhaven, Germany. *e-mail: ljonkers@marum.de

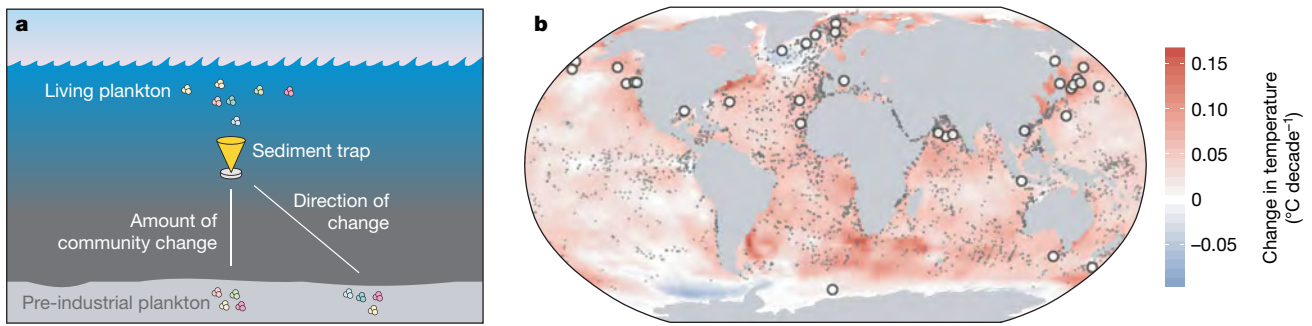


Fig. 1 | Concept of the comparison between Anthropocene and pre-industrial communities. **a**, The integrated living planktonic foraminifera flux from sediment-trap time series is compared to the sedimentary community closest to the trap to quantify community change. The position of the most-similar sedimentary assemblage reveals the direction

of the change. **b**, Grey and white dots indicate the position of the sediment samples and sediment traps, respectively. Background shows the linear sea-surface temperature trend between 1870 and 2015 based on the Hadley Centre Sea Ice and Sea Surface Temperature dataset (HadISST)²⁶.

of environments (open ocean, coastal regions, upwelling regions, and at low and high latitudes), which indicates that species compositions have shifted worldwide. Some parts of the ocean are known to have been cooling during the Anthropocene (Fig. 1). This historical cooling has also affected the composition of Anthropocene assemblages, such that the direction of change inferred from the change in the planktonic foraminifera community is—in the large majority of cases (85%)—consistent with the observed temperature change, irrespective of whether the observed historical trend has been warming or cooling (Fig. 3). The single case in which the species composition indicates warming in a cooling area is in the North Pacific gyre, where the amount of temperature change has been negligible. Thus, we conclude that the chance of finding false warming signatures is low and that the observed pattern in community change is robust, indicating that the communities have responded to the dynamic pattern of changes in sea-surface temperatures induced by global warming. We also observe that time series that show a change in the species community that is inconsistent with historical changes in temperature, are randomly distributed throughout the observational period. Therefore, there is no trend in the degree of consistency, which suggests not only that the faunal composition departed from the pre-industrial baseline but also that it started to do so before the beginning of the sediment-trap observations in ad 1978. If we assume that the inferred community change occurred predominantly after the mid-nineteenth century onset of industrial-era warming, the observed median displacement of approximately 600 km translates into a displacement rate of around 40 km per decade. This is a conservative estimate, because it is likely that the rate of community change accelerated during the twentieth

century and—although there are no comparable data on the rate of community displacement—our estimate is comparable to displacement rates of individual zooplankton species (around 100 km per decade)³.

Even though the pattern of community change in the planktonic foraminifera shows a clear fingerprint of global warming, comparisons between assemblages from the plankton and from the sediment are not straightforward. By using direct observations of integrated annual flux from sediment traps, we obtained a more realistic approximation of sedimentary assemblages than indirect estimates from repeated plankton tows because the effects of ontogeny and seasonal abundance variability can be ruled out. Nevertheless, we also evaluate the effect of other potential biases on our observations, in part because some of the variance in the assemblage change remains unexplained (Fig. 2a). We rule out a temporal sampling bias due to interannual variability as, firstly, the sign of change in individual years is consistent with the long-term signal in 81–92% of the multiyear time series (Fig. 4a) and, secondly, the signal is consistent in 87% of the one-year time series, which should otherwise mostly be affected by noise imposed by interannual variability. The comparisons between modern and pre-industrial assemblages could furthermore be complicated by differential preservation in the sediment^{20,21} and the non-uniform spatial distribution of the sediment samples. We therefore perform sensitivity tests using subsets of the data. We use depth as an indicator of potential preservation bias, as calcite dissolution increases with water depth, and choose 2,000 m as a level below which the assemblages could be affected by dissolution. This is a conservative separation even for the Pacific Ocean, where the lysocline is shallower than in the Atlantic Ocean²². As foraminifera can

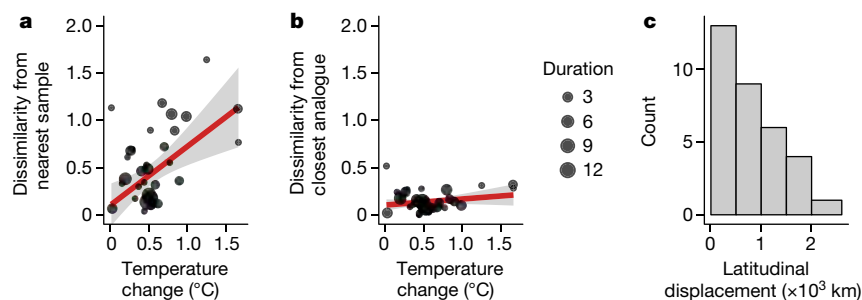


Fig. 2 | Changes in planktonic foraminifera communities in response to Anthropocene sea-surface temperature change. **a**, Dissimilarity to the nearest sediment sample, indicative of the compositional difference between modern and pre-industrial communities, scales with the absolute historical change in temperature at the site of each trap ($n = 33$; $r = 0.53$, weighted by time-series duration, $P = 0.001$; unweighted $r = 0.48$, $P = 0.005$), which suggests that species composition has changed proportionally to temperature change since pre-industrial times. **b**, Dissimilarity between modern and most-similar

pre-industrial species communities, showing that all Anthropocene species communities have more similar analogues elsewhere (difference between **a** and **b**), consistent with a shift of the species composition in the same direction as the temperature change. **a**, **b**, Dots are scaled to the duration of the sediment-trap time series; error envelopes show 95% confidence intervals. **c**, Histogram of latitudinal displacement of planktonic foraminifera communities in kilometres since pre-industrial times (see Methods).

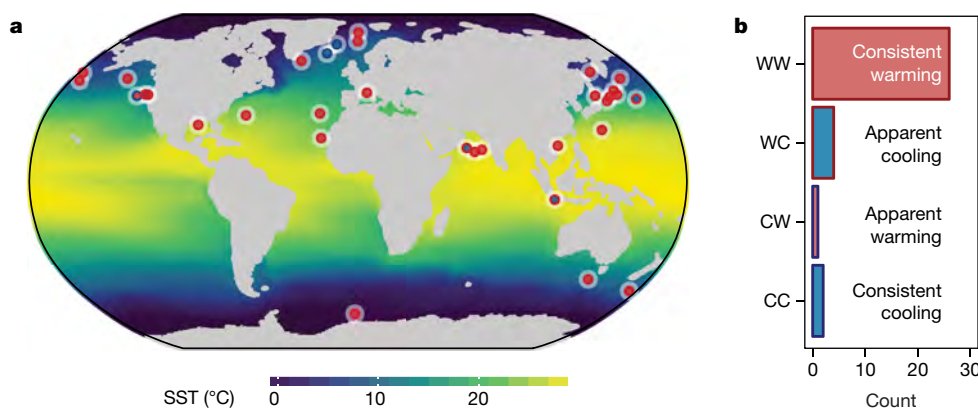


Fig. 3 | Global planktonic foraminifera communities change consistently with historical temperature trends. Planktonic foraminifera community change shows a clear predominance of warming signatures across the oceans, consistent with the known pattern of Anthropocene sea-surface temperature change. **a**, Spatial pattern of the direction of community change (colour fill) and the direction of the historical change in temperature (border colour); both agree in 85% of the observations

($P < 0.001$, two-sided binomial test, $n = 33$). Background colour shading shows mean sea-surface temperature (SST)²⁶ over the observation period (1978–2013). **b**, Histogram of the consistency of Anthropocene community change in response to temperature change. W and C indicate warming and cooling, respectively. The first letter indicates the historical change, and the second indicates change in the species composition.

be transported by ocean currents over hundreds of kilometres during their life cycle and while sinking to the seafloor^{23,24}, we use a 250-km threshold to separate modern samples with a far and nearby pre-industrial counterpart. All four subsets show that in the majority of the cases the direction of change in species communities is consistent with historical temperature change (Fig. 4b). Finally, we also evaluated the sensitivity of our results to the effect of differences in size fractions used to determine the community composition and to the uncertainty in the observational temperature data (see Methods). These tests also confirm that the observed pattern in global planktonic foraminifera community change is robust.

We thus show that there is a globally expressed difference between pre-industrial and Anthropocene marine planktonic foraminifera communities that indicates community turnover, which is in sign and magnitude consistent with global temperature change. As such, we provide evidence in support of observations from shorter time series that lacked a characterization of community state before human influence^{3–5}. The community change in planktonic foraminifera unambiguously shows

that human influence has considerably altered their species communities across the globe. This has important implications for the calibration of palaeoclimate proxies based on planktonic foraminifera, because foraminifera that are preserved in the sediment no longer reflect oceanographic conditions above the site of deposition. We also suggest that the described shifts in planktonic foraminifera are indicative of a more-general phenomenon across marine ecosystems, in which present-day assemblages differ from historical ones in a way that reflects the change in environmental conditions since the onset of the Anthropocene. These findings place emphasis on the recent discussion on how well communities are adapted to rapid environmental change: if the potential for spatial displacement and adaptation lags behind the rate of change in the environment, modern-day assemblages may always show a trait distribution with suboptimal fitness²⁵. Beyond single species, this lag may be especially explicit at the level of communities given the time needed to establish new interaction networks. This, in turn, has potentially large effects on ecosystem functioning as well as on the services that marine ecosystems deliver to society.

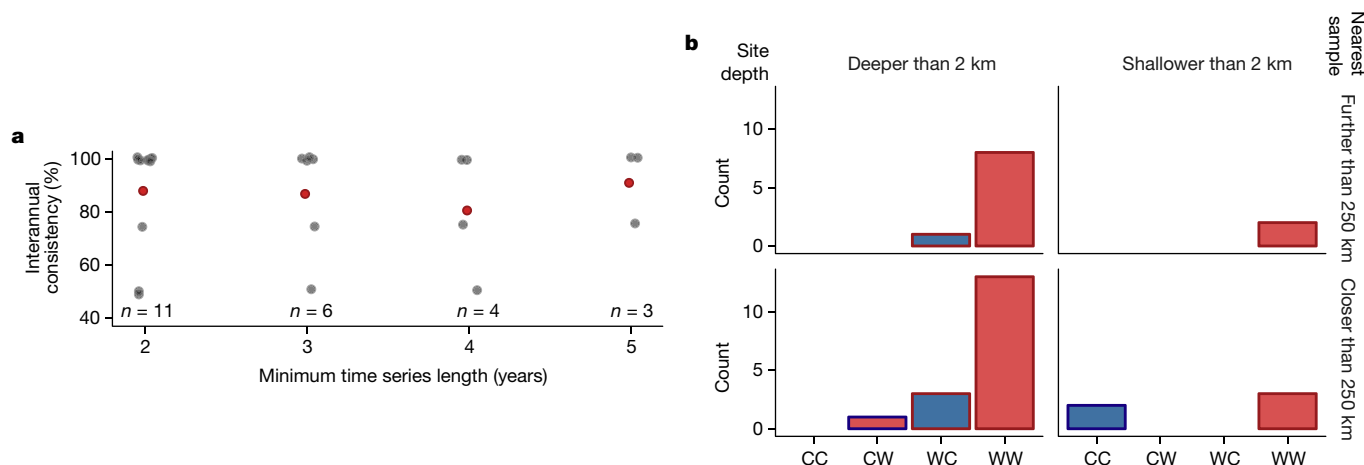


Fig. 4 | Robustness of the sign of change in planktonic foraminifera community composition. **a**, Proportion of individual years in multi-year time series that show a community change consistent with the long-term mean (grey dots, jittered for visibility, show individual time series; red dots show the mean). We restricted this analysis to time series with a duration of at least two years and years without gaps longer than three months. On average, 81–92% of the individual years are consistent with the long-term average and this consistency is robust against time-series length and

imputation. This indicates that our observations are not biased by single anomalous years but instead reflect a robust change in global species composition. **b**, Sensitivity tests show that our results are not dependent on differential preservation and non-uniform spatial distribution of the sediment samples, as each subset is dominated by changes in the species community that are consistent in sign with historical changes in temperature. Labels are as in Fig. 3b.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-019-1230-3>.

Received: 1 August 2018; Accepted: 26 April 2019;

Published online 22 May 2019.

- IPCC. *Climate Change 2013: The Physical Science Basis* (eds Stocker, T. F. et al.) (Cambridge Univ. Press, 2013).
- Abram, N. J. et al. Early onset of industrial-era warming across the oceans and continents. *Nature* **536**, 411–418 (2016).
- Poloczanska, E. S. et al. Global imprint of climate change on marine life. *Nat. Clim. Change* **3**, 919–925 (2013).
- Beaugrand, G., McQuatters-Gollop, A., Edwards, M. & Goberville, E. Long-term responses of North Atlantic calcifying plankton to climate change. *Nat. Clim. Change* **3**, 263–267 (2013).
- Hoegh-Guldberg, O. & Bruno, J. F. The impact of climate change on the world's marine ecosystems. *Science* **328**, 1523–1528 (2010).
- Waters, C. N. et al. The Anthropocene is functionally and stratigraphically distinct from the Holocene. *Science* **351**, aad2622 (2016).
- Field, D. B., Baumgartner, T. R., Charles, C. D., Ferreira-Bartrina, V. & Ohman, M. D. Planktonic foraminifera of the California Current reflect 20th-century warming. *Science* **311**, 63–66 (2006).
- Spielhagen, R. F. et al. Enhanced modern heat transfer to the Arctic by warm Atlantic Water. *Science* **331**, 450–453 (2011).
- Siccha, M. & Kucera, M. ForCenS, a curated database of planktonic foraminifera census counts in marine surface sediment samples. *Sci. Data* **4**, 170109 (2017).
- Rosenzweig, C. et al. Attributing physical and biological impacts to anthropogenic climate change. *Nature* **453**, 353–357 (2008).
- Hillebrand, H. et al. Biodiversity change is uncoupled from species richness trends: consequences for conservation and monitoring. *J. Appl. Ecol.* **55**, 169–184 (2018).
- Gonzalez, A. et al. Estimating local biodiversity change: a critique of papers claiming no net loss of local diversity. *Ecology* **97**, 1949–1960 (2016).
- Morey, A. E., Mix, A. C. & Pisias, N. G. Planktonic foraminiferal assemblages preserved in surface sediments correspond to multiple environment variables. *Quat. Sci. Rev.* **24**, 925–950 (2005).
- Bé, A. W. H. & Tolderlund, D. S. in *The Micropaleontology of Oceans* (eds Funnell, B. M. & Riedel, W. R.) Ch. 6, 105–149 (Cambridge Univ. Press, 1971).
- Morard, R. et al. Surface ocean metabarcoding confirms limited diversity in planktonic foraminifera but reveals unknown hyper-abundant lineages. *Sci. Rep.* **8**, 2539 (2018).
- Rebotim, A. et al. Factors controlling the depth habitat of planktonic foraminifera in the subtropical eastern North Atlantic. *Biogeosciences* **14**, 827–859 (2017).
- CLIMAP Project Members. *Seasonal Reconstruction of the Earth's surface at the Last Glacial Maximum. Map and Chart Series MC-36* (ed. McIntyre, A.) (Geological Society of America, 1981).
- Kucera, M. et al. Reconstruction of sea-surface temperatures from assemblages of planktonic foraminifera: multi-technique approach based on geographically constrained calibration data sets and its application to glacial Atlantic and Pacific Oceans. *Quat. Sci. Rev.* **24**, 951–998 (2005).
- Ruddiman, W. F., Tolderlund, D. S. & Bé, A. W. H. Foraminiferal evidence of a modern warming of the North Atlantic Ocean. *Deep Sea Res.* **17**, 141–155 (1970).
- Berger, W. H. Planktonic Foraminifera: selective solution and paleoclimatic interpretation. *Deep Sea Res.* **15**, 31–43 (1968).
- Berger, W. H. Planktonic Foraminifera: selective solution and the lysocline. *Mar. Geol.* **8**, 111–138 (1970).
- Archer, D. E. An atlas of the distribution of calcium carbonate in sediments of the deep sea. *Glob. Biogeochem. Cycles* **10**, 159–174 (1996).
- von Gyldefeldt, A.-B., Carstens, J. & Meincke, J. Estimation of the catchment area of a sediment trap by means of current meters and foraminiferal tests. *Deep Sea Res.* **47**, 1701–1717 (2000).
- van Sebille, E. et al. Ocean currents generate large footprints in marine palaeoclimate proxies. *Nat. Commun.* **6**, 6521 (2015).
- Enquist, B. J. et al. in *Advances in Ecological Research* Vol. 52 (eds Pawar, S. et al.) 249–318 (Academic, 2015).
- Rayner, N. A. et al. Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. *J. Geophys. Res.* **108**, 4407 (2003).

Acknowledgements We thank R. Reuter for help with foraminifera analysis and acknowledge funding by the Volkswagen Stiftung for the MarBAS (Marine Biodiversität—Analyse über zeitliche und räumliche Skalen) project as well as by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) through Germany's Excellence Strategy (EXC-2077, grant no 390741603). M.K. was funded through DFG-Research Center/Cluster of Excellence 'The Ocean in the Earth System'.

Reviewer information Nature thanks Andrew J. Fraass, Anthony Richardson and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions L.J. and M.K. designed research. L.J. compiled and analysed the data. All authors discussed the results and contributed to the writing of the manuscript.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-019-1230-3>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-019-1230-3>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to L.J.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

METHODS

Data. Quality-controlled and taxonomically harmonized core top assemblage data were obtained from the ForCenS dataset⁹. Samples for which species groups and species forms (for example, differently coiled morphospecies) were not resolved were excluded ($n = 128$) and in cases in which multiple samples were available for the same location ($n = 303$), one was randomly selected. The modern assemblages are derived by integrating shell flux time series from moored sediment traps (Extended Data Table 1). Sediment-trap time series selection criteria followed a previous study²⁷ with the additional constraint that only time series with full taxonomic resolution (considering the same range of species as in the sediment samples) were included. The taxonomy of the time series was harmonized with that of the core top data following previously published criteria⁹. To eliminate the effect of seasonality in the flux when deriving the flux-weighted assemblages, only time series with a length of at least 345 days were analysed. Gaps in the time series were linearly interpolated or, where possible, filled using a smoothed version (LOESS, span 0.1) of the multiyear median flux pattern. To make full use of the available data and to estimate the influence of interannual flux variability, annual assemblages were calculated for each year for which at least three months of data were available, using the smoothed flux pattern to impute missing values. As is common in palaeo-ecological studies, the dissimilarity between the pre-industrial and modern species assemblages was assessed using square chord distance between species relative abundances because this metric has been shown to be most effective in identifying analogues in microfossil datasets²⁸.

Various planktonic foraminifera morphospecies consist of multiple, genetically different, cryptic species²⁹. To minimize the effect of differences in the ecology of cryptic species, dissimilarity was determined regionally¹⁸ such that sediment samples with the most-similar assemblages were sought only within the same oceanic basin as the analysed sediment-trap time series.

The latitudinal displacement of the species communities was estimated from the dissimilarity between the modern assemblage and the nearest sediment assemblage. We used linear relationships between dissimilarity and the latitudinal component of distance derived from the sediment data. To account for regional variability in spatial trends in biodiversity, we established these relationships for each sediment trap site, using the latitudinal distance from the nearest sediment sample within a radius of 5,000 km. Regression models were forced through the origin (because at zero distance, the assemblages should be identical) and established for the first 2,500 km latitudinal distance, as this is the distance at which dissimilarity tends to saturate and the distance–dissimilarity relationship changes slope (see example in Extended Data Fig. 2).

Sea-surface temperature data are from the HadISST dataset²⁶. We use data from the 1870–1899 period within a 100-km radius around each sampling location to derive an estimate of the mean temperature. The direction of historical change in temperature for each sediment-trap location was derived from the difference between the mean temperature at the time of sediment-trap sample collection and the late-nineteenth-century mean. This temperature change provides an estimate of the temperature change at each site that is indicative of warming or cooling since approximately the beginning of industrial-era warming. However, given the uncertainty in the age of the sedimentary species assemblages, it is only an approximation of the temperature change since the time of deposition of the sediments. We nevertheless consider this adequate for the purpose of our analysis.

All analyses were done in R³⁰ using packages rioja³¹, reshape2³², ggplot2³³, geosphere³⁴, readxl³⁵, Hmisc³⁶, raster³⁷, sp^{38,39} and rgdal⁴⁰.

Pre-industrial age of core tops. Marine-sediment archives provide a temporally integrated record of species assemblages. The length of this record depends on the sampling resolution, the sediment accumulation rate and the depth of the layer mixed by bioturbation. The mean age of the core top samples can thus be estimated using the depth solution of a previously published study⁴¹ and reasonable approximations of the sediment accumulation rate and bioturbation depth. Using an empirical relationship between water depth and Holocene sediment accumulation rate⁴², we calculate a median sediment accumulation rate of the core top samples of 5.9 (interquartile range, 3.8–6.2) cm per thousand years. For bioturbation depth, we use a global average⁴³ of 9.8 ± 4.5 cm. These analyses yield estimated mean ages of the core top sediments of centuries to millennia, warranting their use as an integrated pre-industrial baseline of the planktonic foraminifera assemblages (Extended Data Fig. 1). Note that Extended Data Fig. 1 provides the average age of non-normal distribution of the ages of all sedimentary particles (including foraminifera), such that all sediments also contain foraminifera within the age range of the temperature estimates before marked human influences.

Effect of shell size. Planktonic foraminifera differ in their mean size among species and species assemblages may therefore vary as a function of the analysed size fraction. Whereas the core top samples are all $> 150 \mu\text{m}$, our shell flux time series compilation also includes data for sizes $> 125 \mu\text{m}$. Even though cold-water species are more abundant in smaller size fractions⁴⁴ and inclusion of data for sizes of

$< 150 \mu\text{m}$ would thus bias our results towards cooling rather than warming we nevertheless assessed the influence of size fraction in two ways.

First, we included an additional sensitivity test for the consistency of community change (analogous to Fig. 4) separating the time series by size fraction (Extended Data Fig. 3). Both groups show a change in the community composition consistent with temperature change in the large majority of the cases, which indicates that our results are insensitive to the size fraction of the foraminifera in the sediment-trap time series.

Second, we carried out separate analyses of the coarse- and fine-fraction data from the seven time series for which size-fractionated data are available (CAS, CCG, CCM, CCN, EAS, MBL and WAS; see Extended Data Table 1 for details). In 6 out of these 7 cases, the direction of community change indicated by the assemblages $> 125 \mu\text{m}$ is identical to the change estimate from the assemblages $> 150 \mu\text{m}$, or indicative of cooling. This means that inclusion of data from time series with the small-size fraction data are much more likely to suggest cooling than warming.

Importantly, the two cases of a consistent shift towards cooler assemblages both pertain to data from the $> 150\text{-}\mu\text{m}$ fraction (Extended Data Table 1) and hence do not reflect size-related biases. The pattern of assemblage change is therefore a robust, if conservative, estimate of the change in planktonic foraminifera community change.

Effect of choice of temperature dataset. Historical sea-surface temperature data that predate the onset of large-scale shipboard measurements and satellite observations are associated with uncertainty, which is reflected in the differences between different data products. To evaluate the effect of the choice of temperature data product, we also conducted our analysis using ERSST version 5 data⁴⁵. Compared to the HadISST data, ERSST data have a coarse resolution ($2^\circ \times 2^\circ$ compared to $1 \times 1^\circ$), but the data extend back to ad 1854, offering the possibility to obtain a better estimate of industrial-era warming. The ERSST v.5 data (also integrated over 30 years, but shifted backwards to ad 1854 to make full use of the longer temporal extent of the ERSST data) show generally less temperature change compared to HadISST⁴⁶. However, our results are largely insensitive to the data product used and the two sea-surface temperature products reveal the same patterns. The scaling between dissimilarity and temperature change is similar, albeit with a larger uncertainty for ERSST (Extended Data Fig. 4a), and the changes in species community are also largely consistent with the temperature change estimated from the ERSST data (Extended Data Fig. 4b). This makes us confident that the observed patterns in the change in species communities are robust.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

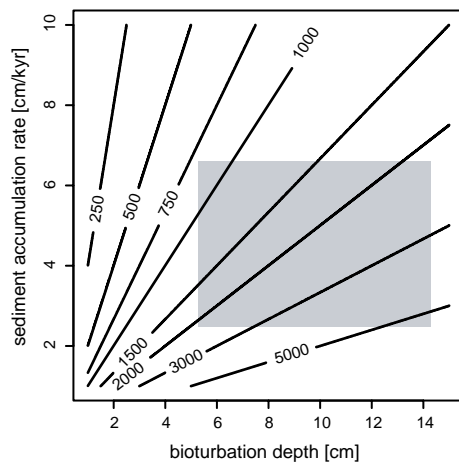
The ForCenS core top planktonic foraminifera dataset is available at Pangaea (<https://doi.org/10.1594/PANGAEA.873570>) and the HadISST data are available from the UK Met Office (<https://www.metoffice.gov.uk/hadobs/hadisst/>). NOAA ERSST v.5 data were provided by the NOAA/OAR/ESRL PSD (<https://www.esrl.noaa.gov/psd/>). Taxonomically harmonized shell flux data are available at <https://doi.org/10.5281/zenodo.2638013>.

Code availability

Code is available at <https://doi.org/10.5281/zenodo.2638013>.

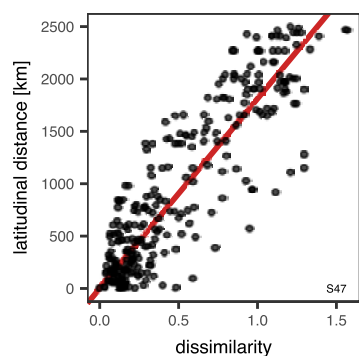
- Jonkers, L. & Kučera, M. Global analysis of seasonality in the shell flux of extant planktonic Foraminifera. *Biogeosciences* **12**, 2207–2226 (2015).
- Prell, W. *The Stability of Low-Latitude Sea-Surface Temperatures, an Evaluation of the CLIMAP Reconstruction with Emphasis on the Positive SST Anomalies*. Report No. TR025 (US Department of Energy, 1985).
- Darling, K. F. & Wade, C. M. The genetic diversity of planktic foraminifera and the global distribution of ribosomal RNA genotypes. *Mar. Micropaleontology* **67**, 216–238 (2008).
- R Core Team. *R: A Language and Environment for Statistical Computing*. <https://www.R-project.org/> (R Foundation for Statistical Computing, 2016).
- Juggins, S. *rioja: Analysis of Quaternary Science Data*. R package version 0.9-15.1 <http://cran.r-project.org/package=rioja> (2017).
- Wickham, H. Reshaping data with the reshape package. *J. Stat. Softw.* **21**, 1–20 (2007).
- Wickham, H. *ggplot2: Elegant Graphics for Data Analysis* (Springer, 2016).
- Hijmans, R. J., Williams, E. & Vennes, C. geosphere: Spherical Trigonometry. R package version 1.5-7 <https://CRAN.R-project.org/package=geosphere> (2017).
- Wickham, H. & Bryan, J. readxl: Read Excel Files. R package version 1.1.0 <https://CRAN.R-project.org/package=readxl> (2018).
- Harrell, F. E. Jr. Hmisc: Harrell Miscellaneous. R package version 4.1-1 <https://CRAN.R-project.org/package=Hmisc> (2018).
- Hijmans, R. J. et al. raster: Geographic Data Analysis and Modeling. R package version 2.6-7. <https://CRAN.R-project.org/package=raster> (2017).
- Pebesma, E. J. & Bivand, R. S. Classes and methods for spatial data in R. *R News* **5**, 9–13 (2005).

39. Bivand, R. S., Pebesma, E. & Gómez-Rubio, V. *Applied Spatial Data Analysis with R* (Springer, 2008).
40. Bivand, R. et al. rgdal: Bindings for the 'Geospatial' Data Abstraction Library. R package version 1.3-1 <https://CRAN.R-project.org/package=rgdal> (2018).
41. Berger, W. H. & Heath, G. R. Vertical mixing in pelagic sediments. *J. Mar. Res.* **26**, 134–143 (1968).
42. Burwicz, E. B., Rüpke, L. H. & Wallmann, K. Estimation of the global amount of submarine gas hydrates formed via microbial methane formation based on numerical reaction-transport modeling and a novel parameterization of Holocene sedimentation. *Geochim. Cosmochim. Acta* **75**, 4562–4576 (2011).
43. Boudreau, B. P. Mean mixed depth of sediments: the wherefore and the why. *Limnol. Oceanogr.* **43**, 524–526 (1998).
44. Al-Sabouni, N., Kucera, M. & Schmidt, D. N. Vertical niche separation control of diversity and size disparity in planktonic foraminifera. *Mar. Micropaleontol.* **63**, 75–90 (2007).
45. Huang, B. et al. NOAA Extended Reconstructed Sea Surface Temperature (ERSST). Version 5 <https://doi.org/10.7289/V5T72FNM> (NOAA National Centers for Environmental Information, 2017).
46. Huang, B. et al. Further exploring and quantifying uncertainties for extended reconstructed sea surface temperature (ERSST) version 4 (v4). *J. Clim.* **29**, 3119–3142 (2016).
47. Asahi, H. & Takahashi, K. A 9-year time-series of planktonic foraminifer fluxes and environmental change in the Bering Sea and the central subarctic Pacific Ocean, 1990–1999. *Prog. Oceanogr.* **72**, 343–363 (2007).
48. Deuser, W. G. & Ross, E. H. Seasonally abundant planktonic foraminifera of the Sargasso Sea; succession, deep-water fluxes, isotopic compositions, and paleoceanographic implications. *J. Foraminiferal Res.* **19**, 268–293 (1989).
49. Deuser, W. G., Ross, E. H., Hemleben, C. & Spindler, M. Seasonal changes in species composition, numbers, mass, size, and isotopic composition of planktonic foraminifera settling into the deep Sargasso Sea. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **33**, 103–127 (1981).
50. Northcote, L. C. & Neil, H. L. Seasonal variations in foraminiferal flux in the Southern Ocean, Campbell Plateau, New Zealand. *Mar. Micropaleontol.* **56**, 122–137 (2005).
51. Gupta, M. V. S., Curry, W. B., Ittekkot, V. & Muralinath, A. S. Seasonal variation in the flux of planktic Foraminifera; sediment trap results from the Bay of Bengal, northern Indian Ocean. *J. Foraminiferal Res.* **27**, 5–19 (1997).
52. Žarić, S., Donner, B., Fischer, G., Mulitza, S. & Wefer, G. Sensitivity of planktic foraminifera to sea surface temperature and export production as derived from sediment trap data. *Mar. Micropaleontol.* **55**, 75–105 (2005).
53. Reuter, R. T., Jonkers, L. & Kucera, M. Planktonic foraminifera shell flux data from sediment trap CB-3. PANGAEA <https://doi.org/10.1594/PANGAEA.899732> (2016).
54. Ortiz, J. D. & Mix, A. C. The spatial distribution and seasonal succession of planktonic foraminifera in the California Current off Oregon, September 1987 – September 1988. *Geol. Soc. Lond. Spec. Publ.* **64**, 197–213 (1992).
55. Jensen, S. *Planktische Foraminiferen im Europäischen Nordmeer: Verbreitung und Vertikalfuß sowie ihre Entwicklung während der letzten 15000 Jahre*. PhD thesis, Univ. Kiel (1998).
56. Poore, R. Z., Tedesco, K. A. & Spear, J. W. Seasonal flux and assemblage composition of planktic foraminifera from a sediment-trap study in the northern Gulf of Mexico. *J. Coast. Res.* **63**, 6–19 (2013).
57. Reynolds, C. E., Richey, J. N. & Poore, R. Z. *Seasonal Flux and Assemblage Composition of Planktic Foraminifera from the Northern Gulf of Mexico, 2008–2012*. US Geological Survey Open-File Report 2013–1243 <https://doi.org/10.3133/ofr20131243> (USGS, 2013).
58. Jonkers, L., Reynolds, C. E., Richey, J. & Hall, I. R. Lunar periodicity in the shell flux of planktonic foraminifera in the Gulf of Mexico. *Biogeosciences* **12**, 3061–3070 (2015).
59. Wolfteich, C. M. *Satellite-Derived Sea Surface Temperature, Mesoscale Variability, And Foraminiferal Production in the North Atlantic*. MSc thesis, MIT and WHOI (1994).
60. Jonkers, L., Brummer, G.-J. A., Peeters, F. J. C., van Aken, H. M. & De Jong, M. F. Seasonal stratification, shell flux, and oxygen isotope dynamics of left-coiling *N. pachyderma* and *T. quinqueloba* in the western subpolar North Atlantic. *Paleoceanography* **25**, PA2204 (2010).
61. Jonkers, L., van Heuven, S., Zahn, R. & Peeters, F. J. C. Seasonal patterns of shell flux, $\delta^{18}\text{O}$ and $\delta^{13}\text{C}$ of small and large *N. pachyderma* (s) and *G. bulloides* in the subpolar North Atlantic. *Paleoceanography* **28**, 164–174 (2013).
62. Reuter, R. T., Jonkers, L., Brummer, G. J. & Kucera, M. Planktonic foraminifera shell flux data from sediment trap IRM-1. PANGAEA <https://doi.org/10.1594/PANGAEA.899733> (2018).
63. Mohtadi, M. et al. Low-latitude control on seasonal and interannual changes in planktonic foraminiferal flux and shell geochemistry off south Java: A sediment trap study. *Paleoceanography* **24**, PA1201 (2009).
64. Rigual-Hernández, A. S., Sierro, F. J., Bárcena, M. A., Flores, J. A. & Heussner, S. Seasonal and interannual changes of planktic foraminiferal fluxes in the Gulf of Lions (NW Mediterranean) and their implications for paleoceanographic studies: two 12-year sediment trap records. *Deep Sea Res.* **66**, 26–40 (2012).
65. Donner, B. & Wefer, G. Flux and stable isotope composition of *Neoglobobulimina pachyderma* and other planktonic foraminifera in the Southern Ocean (Atlantic sector). *Deep Sea Res.* **41**, 1733–1743 (1994).
66. Storz, D., Schulz, H., Wanek, J. J., Schulz-Bull, D. E. & Kučera, M. Seasonal and interannual variability of the planktic foraminiferal flux in the vicinity of the Azores Current. *Deep Sea Res.* **56**, 107–124 (2009).
67. Kuroyanagi, A., Kawahata, H., Nishi, H. & Honda, M. C. Seasonal changes in planktonic foraminifera in the northwestern North Pacific Ocean: sediment trap experiments from subarctic and subtropical gyres. *Deep Sea Res.* **49**, 5627–5645 (2002).
68. Sagawa, T., Kuroyanagi, A., Irino, T., Kuwae, M. & Kawahata, H. Seasonal variations in planktonic foraminiferal flux and oxygen isotopic composition in the western North Pacific: implications for paleoceanographic reconstruction. *Mar. Micropaleontol.* **100**, 11–20 (2013).
69. Alderman, S. E. *Planktonic Foraminifera in the Sea of Okhotsk: Population and Stable Isotopic Analysis from a Sediment Trap*. MSc thesis, MIT and WHOI (1996).
70. Sautter, L. R. & Thunell, R. C. Seasonal succession of planktonic foraminifera; results from a four-year time-series sediment trap experiment in the Northeast Pacific. *J. Foraminiferal Res.* **19**, 253–267 (1989).
71. King, A. L. & Howard, W. R. Planktonic foraminiferal flux seasonality in Subantarctic sediment traps: a test for paleoclimate reconstructions. *Paleoceanography* **18**, 1019 (2003).
72. Curry, W. B., Ostermann, D. R., Gupta, M. V. S. & Ittekkot, V. Foraminiferal production and monsoonal upwelling in the Arabian Sea: evidence from sediment traps. *Geol. Soc. Lond. Spec. Publ.* **64**, 93–106 (1992).
73. Mohiuddin, M. M., Nishimura, A., Tanaka, Y. & Shimamoto, A. Regional and interannual productivity of biogenic components and planktonic foraminiferal fluxes in the northwestern Pacific Basin. *Mar. Micropaleontol.* **45**, 57–82 (2002).
74. Mohiuddin, M. M., Nishimura, A. & Tanaka, Y. Seasonal succession, vertical distribution, and dissolution of planktonic foraminifera along the Subarctic Front: implications for paleoceanographic reconstruction in the northwestern Pacific. *Mar. Micropaleontol.* **55**, 129–156 (2005).
75. Xiang, R. et al. Seasonal flux variability of planktonic foraminifera during 2009–2011 in a sediment trap from Xisha Trough, South China Sea. *Aquat. Ecosyst. Health Manage.* **18**, 403–413 (2015).

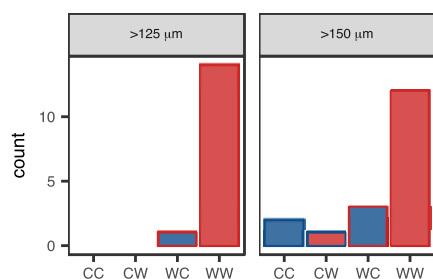


Extended Data Fig. 1 | Pre-industrial age of the sedimentary samples.

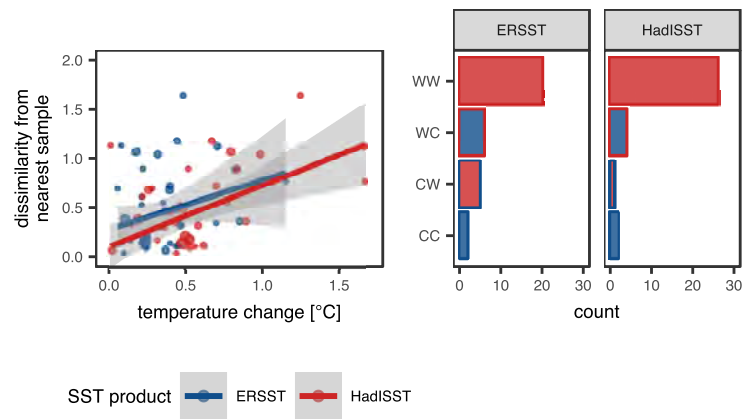
Mean age in years of core top sediment estimate using the depth solution of a previous study⁴¹ (contours). The grey box denotes the likely average ages of the core top sediments based on our best estimate of sediment accumulation rate (in cm per 1,000 years (kyr)) and bioturbation depth. Irrespective of the sampling date (mostly pre-1980), the average sedimentary species composition predates the Anthropocene.



Extended Data Fig. 2 | Linear regression between dissimilarity and latitudinal distance in the sedimentary species assemblages. The relationship (shown in red) is used to estimate the latitudinal displacement based on the dissimilarity between the modern and pre-industrial species composition. Example for time series S47 from the south of New Zealand (Extended Data Table 1).



Extended Data Fig. 3 | Insensitivity of planktonic foraminifera assemblage change to size fraction. The direction of change for planktonic foraminifera species communities (warming or cooling) was inferred from sediment-trap time series for which the samples were larger than 125 μm and larger than 150 μm . Colours and symbols are as in Fig. 3b. W and C indicate warming and cooling, respectively, with the first letter indicating the historical change and the second the change as indicated by the species composition. Both small and large shell sizes are dominated by a change in the species community that is consistent with the direction of historical change in temperatures. The observed pattern is thus insensitive to the inclusion of sediment-trap time series that used a slightly smaller size fraction than the sediment samples.



Extended Data Fig. 4 | Assessing uncertainty in the historical change in temperatures by comparing the HadISST and ERSST temperature products. a, Comparison of the relationships between the historical change in temperature and the difference between the modern and sedimentary species composition (based on linear regression weighted to the duration of the time series; see also Fig. 2a). The relationship has a similar slope for both sea-surface temperature products, even though the

relationship based on ERSST data has a larger uncertainty. Shaded error envelopes show 95% confidence intervals of the regression. **b,** Histograms of consistency and direction of changes in the species communities (Fig. 3a). The pattern of change is broadly similar for both products, which indicates that although the observations are to some degree sensitive to the uncertainty in the historical change in temperatures, they are largely consistent between the two datasets.

Extended Data Table 1 | Sediment-trap time series used to determine modern species compositions

Name	Longitude [east]	Latitude [north]	Water depth [m]	Duration [days]	Minimum size [μm]	Reference	Change*
ABP	-177.00	53.05	3788	3265	125	47	WW
BAT	-64.25	32.08	4200	2233	125	48, 49	WW
CAP	174.15	-52.62	2580	426	150	50	WW
CAS	64.75	14.47	3900	503	150	51	WW
CBL	-20.69	21.15	4150	389	150	52, 53	WW
CCG	-132.02	41.54	3664	364	150	54	CW
CCM	-127.58	42.19	2830	352	150	54	WC
CCN	-125.77	42.09	2829	360	150	54	WW
EAS	68.75	15.47	3770	528	150	51	WW
GLS	0.00	75.00	3720	346	125	55	WW
GOM	-90.30	27.50	1300	431	150	56-58	WW
ILP	-16.00	68.00	1231	352	150	59	CC
IRM	-38.5	59.00	3000	380	150	60-62	WW
JAM	108.00	-8.25	3060	984	150	63	WC
LPL	5.18	43.02	1000	4459	150	64	WW
MAU	-2.50	-64.90	5023	1133	125	65	WW
MBL	-22.00	33.00	5500	762	150	66	WW
MRI	-22.00	63.00	833	352	150	59	CC
NBA	-0.47	69.69	3270	444	125	55	WW
NP4	165.00	40.00	5483	953	125	67	WC
NP5	165.00	50.00	5570	1288	125	67	WW
NPK	155.00	44.00	5370	894	125	67	WW
NWP	141.87	41.56	970	361	125	68	WW
OKH	149.85	53.32	1166	365	150	69	WW
PAP	-145.00	50.00	4256	1438	125	70	WW
S47	142.00	-46.80	4571	494	150	71	WW
SAP	-174.00	49.00	5406	3268	125	47	WW
WAS	60.47	16.32	4015	529	150	72	WC
WC1	137.00	25.00	5100	613	125	73	WW
WC2	147.00	39.00	5335	629	125	73	WW
WC5	150.00	40.97	5615	358	125	74	WW
WC6	155.24	42.01	5578	382	125	74	WW
XTS	110.87	17.43	1694	790	154	75	WW

Data were obtained from previous studies⁴⁷⁻⁷⁵.

*The first letter refers to the historical temperature change, and the second letter refers to the change indicated by the change in the species composition. W and C indicate warming and cooling, respectively.

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- ☐ ☒ The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- ☒ ☐ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- ☐ ☒ The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- ☒ ☐ A description of all covariates tested
- ☒ ☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- ☐ ☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- ☐ ☒ For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
Give P values as exact values whenever suitable.
- ☒ ☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- ☒ ☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- ☐ ☒ Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection

no software used for collection

Data analysis

R version 3.5; publicly available packages listed in methods

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The ForCens core top planktonic foraminifera data set is available at Pangaea (<https://doi.pangaea.de/10.1594/PANGAEA.873570>) and the HadISST data are available from the UK Met Office (<https://www.metoffice.gov.uk/hadobs/hadisst/>). NOAA ERSST V5 data were provided by the NOAA/OAR/ESRL PSD, Boulder, Colorado, USA, from their Web site at <https://www.esrl.noaa.gov/psd/>. Code and taxonomically harmonised shell flux data will be made available on github/zenodo.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☒ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	We compare species communities (relative abundances) of fossil and modern marine microplankton (planktonic foraminifera) using commonly used dissimilarity metrics. Data on the modern species communities is derived from sediment traps and fluxes have been integrated to allow comparison with the fossil data. The approach is described in the method section.
Research sample	We used published data only; fossil species community data from Siccha, M., and Kucera, M.: ForCenS, a curated database of planktonic foraminifera census counts in marine surface sediment samples, Scientific Data, 4, 170109, 10.1038/sdata.2017.109, 2017. We chose this dataset as it is the most extensive and most recent compilation to date. Sediment trap data were compiled from literature (see Jonkers, L., and Kučera, M.: Global analysis of seasonality in the shell flux of extant planktonic Foraminifera, Biogeosciences, 12, 2207-2226, 10.5194/bg-12-2207-2015, 2015.). We have updated this compilation in order to include all (to the best of our knowledge, see also reviews) sediment trap time series. A full list is provided in the method section of the manuscript
Sampling strategy	We used compiled literature data
Data collection	We used previously compiled and published data, collection methods are described in the relevant publications listed in the method section
Timing and spatial scale	We used previously published data. The fossil data are from surface deep sea sediments sampled over the past decades (age assessed in manuscript) and the modern data from moored sediment traps are from over 30 different studies since 1978. Both data sets are global in their extent
Data exclusions	Criteria for inclusion in our shell flux compilation were predetermined and 1) a minimum time series length of 345 days in order to reduce the effect of seasonality; 2) complete taxonomic resolution in order to sensibly compare plankton and sediment species assemblages and 3) no indication of resuspension (i.e. the presence of benthic foraminifera) to make sure that we are looking at a primary signal.
Reproducibility	no experiments conducted
Randomization	no experiments conducted; not applicable to observational time series from sediment traps used here
Blinding	no experiments conducted; not applicable to observational time series from sediment traps used here
Did the study involve field work?	<input type="checkbox"/> Yes <input checked="" type="checkbox"/> No

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data

Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

Mating preferences of selfish sex chromosomes

Pavitra Muralidhar^{1,2*}

The evolution of female mating preferences for harmful male traits is a central paradox of sexual selection^{1–9}. Two dominant explanations for this paradox^{8,10} are Fisher’s runaway process, which is based on genetic correlations between preference and trait^{1,3,4}, and Zahavi’s handicap principle, in which the trait is an honest costly signal of male quality^{2,6,8,11}. However, both of these explanations require the exogenous initial spread of female preferences before harmful male traits can evolve^{1–4,6,8,11}. Here I present a mechanism for the evolution of female mating preferences for harmful male traits that is based on the selfish evolutionary interests of sex chromosomes. I demonstrate that female-biased genetic elements—such as the W and X sex chromosomes—will evolve mating preferences for males who display traits that reduce their fitness and/or that of their male offspring, but increase fitness in female offspring. In particular, W-linked preferences can cause nearly lethal male traits to sweep to fixation. Sex-linked preferences can drive the evolution of traits such as ornamental handicaps and male parental care, and can explain variation in ornamentation and behaviour across taxa with divergent sex-determining mechanisms.

Female mating preferences should evolve to maximize total offspring fitness⁷. Intra-genomic conflict complicates this picture, because females can carry multiple genetic elements that have sex-biased transmission^{12,13}. This is clearest for the W chromosome in female-heterogametic (ZW) species, such as birds: autosomes spend as many generations in males as in females, but the W chromosome is only ever carried by females^{12–14}. A preference encoded on the W chromosome should therefore evolve to maximize the total fitness of daughters, with no regard for the fitness of sons (to whom it is not transmitted).

Traits that increase the fitness of daughters at the expense of the fitness of fathers or sons can take many forms. One major category is sexually antagonistic traits, which increase fitness in one sex but reduce it in the other^{13,15}. Such traits are common in natural populations^{16–19}. Usually, to avoid elimination by natural selection, a sexually antagonistic trait must either confer a fitness advantage when averaged across the sexes or be sex-linked^{13,15}. In previously studied scenarios, these conditions limit the fitness cost that can be imposed on the sex for which the trait is deleterious; here, I show that this is not true when mating preferences for sexually antagonistic traits are encoded on a sex chromosome.

Previous theoretical work has separately considered the roles of sexual antagonism^{9,20}, sex linkage²¹, sex determination^{22,23} and reinforcing female preferences^{5,20} in sexual selection. However, to my knowledge, no previous model has examined the co-evolution of sex-linked female preferences for autosomal, sexually antagonistic traits.

To examine this process, I considered a two-locus population genetic model of a ZW species, with an autosomal ‘trait’ locus and a W-linked ‘preference’ locus (for full details, see Methods). Z-linked and X-linked preferences are discussed below. In this model, two alleles segregate at the trait locus: the wild-type allele (*t*) and the mutant allele (*T*), which increases female viability (by a factor $1 + s_f$ for *TT* homozygotes and $1 + h_T s_f$ for *Tt* heterozygotes) but reduces male viability (by $1 - s_m$ for *TT* and $1 - h_T s_m$ for *Tt*). s_f is the strength of the viability advantage of the *T* allele in females; s_m is the strength of its viability disadvantage in males. h_T is the dominance of the *T* allele with respect to the *t* allele.

The alleles mutate from one to the other at a symmetrical rate u per replication. I assume that $s_m > s_f$, so that *T* is selected against in the absence of other forces. Two alleles segregate at the W-linked preference locus: the wild-type allele *p* and the mutant allele *P*, the bearers of which (always female) have a greater propensity to mate with trait-expressing males (by a factor $\alpha > 1$ for *TT* males and α^{h_T} for *Tt* males, where α is the strength of the preference). Here I assume $h_T = 1/2$ (co-dominance), although the qualitative features of the results do not depend on this assumption (see Extended Data Fig. 1).

It can be proven (Supplementary Information) that the *P* allele increases in frequency as long as the trait locus is polymorphic. Therefore, the *P* allele will fix if there is a source of persistent trait polymorphism, such as recurrent mutation or migration from a population with reduced selection against the trait (Fig. 1). This positive selection arises indirectly. The *P* allele generates a positive genetic correlation between itself and the *T* allele by inducing its bearers to preferentially mate with males that bear the *T* allele. Because the *T* allele increases fitness in females (and the *P* allele is present only in females), this positive association causes the frequency of the *P* allele to rise.

The strength of positive selection acting on the *P* allele depends on several factors. For example, it increases with the strength of the preference induced by the *P* allele, and with the fitness advantage conferred by the *T* allele in females. To investigate the strength of selection in favour of the *P* allele, I compared the strengths observed in several configurations of the model to those observed in the standard two-locus autosomal model of Fisherian sexual selection⁴. In this model, selection for low-frequency W-linked preferences is consistently stronger—often by orders of magnitude—than selection for analogous autosomal preferences, even when the latter start at the high frequencies required for the trait to spread (Supplementary Information).

Selection on the *T* allele depends on its cost to males, its benefit to females and the proportion of females that carry the *P* allele. If the strength of the preference is sufficiently large ($\alpha \gtrsim 1/[(1 - s_m)(1 + s_f)]$, Supplementary Information), selection favours the *T* allele for frequencies of the *P* allele above a certain threshold. Because the *P* allele inevitably rises to fixation, this threshold is eventually exceeded and the *T* allele spreads. The resultant equilibrium is one in which many males exhibit a trait that severely impairs their survival, and all females exhibit a strong mating preference for these low-viability males (Fig. 1, Extended Data Fig. 2b). This can occur even for traits that are nearly lethal to males but that confer only a small advantage to females (Fig. 1d). If $\alpha \lesssim 1/[(1 - s_m)(1 + s_f)]$ instead, the *T* allele remains at low frequency, even after the *P* allele has fixed. In this equilibrium, all females prefer low-viability males despite these males being nearly absent from the population (Fig. 1, Extended Data Fig. 2a).

In this model, the spread of the harmful male trait does not require initial neutral drift of—or exogenous selection for—the mutant preference, unlike in analogous two-locus models of Fisher’s runaway process^{4,8} and Zahavi’s handicap model^{6,8,11}. By extension, preferences that impose fitness costs on females (for example, by reducing their probability of finding a mate) can invade from low frequency in this model, unlike in comparable major-effect runaway and handicap models (which are very sensitive to costs of female preferences⁸).

¹Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA, USA. ²Program for Evolutionary Dynamics, Harvard University, Cambridge, MA, USA. e-mail: pmuralidhar@g.harvard.edu

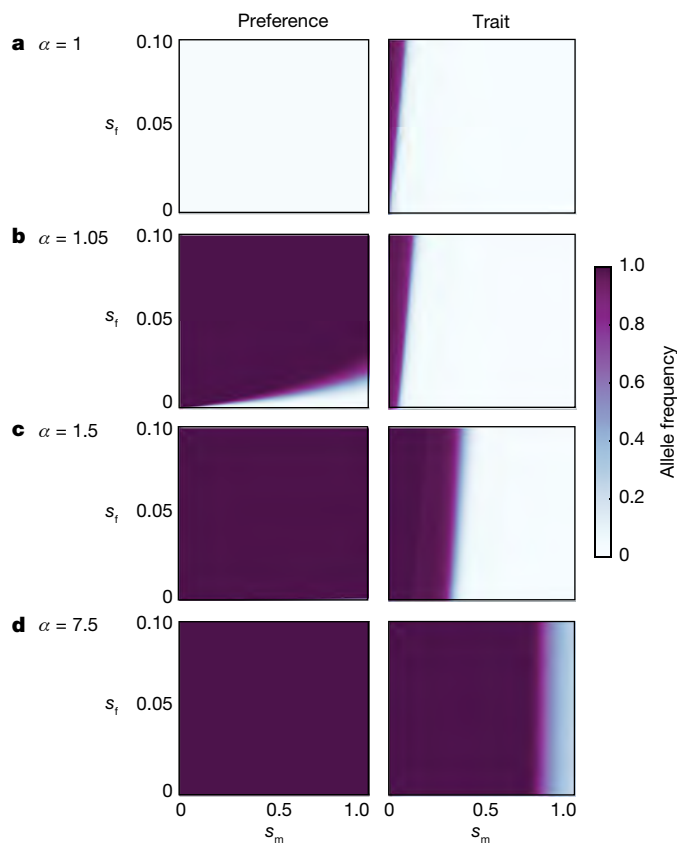


Fig. 1 | Evolution of W-linked preferences for sexually antagonistic traits. Long-run frequencies of the W-linked *P* allele and the autosomal sexually antagonistic *T* allele after 5×10^6 generations, each having started at 1% frequency. The *T* and *t* alleles mutate from one to the other at a rate of 10^{-3} per replication. **a**, When the *P* allele induces no preference ($\alpha = 1$), the sexually antagonistic *T* allele reaches high frequency only when it increases viability on average across the sexes (that is, when $(1 + s_f)(1 - s_m) > 1$). **b**, Even when the preference encoded by the *P* allele is weak ($\alpha = 1.05$), the *P* allele is positively selected for, and fixes in a large region of the parameter space in which the sex-averaged viability effect of the *T* allele is negative (that is, where $(1 + s_f)(1 - s_m) < 1$). Fixation of the *P* allele pushes the *T* allele to high frequency over a small region of parameter space, in which the cost of the trait to males (s_m) is not too large compared to the benefit of the trait to females (s_f). **c**, For slightly higher strengths of the preference encoded by the *P* allele ($\alpha = 1.5$), the allele always fixes and the *T* allele attains high frequency in regions of parameter space where male costs are very high. **d**, When the preference is strong ($\alpha = 7.5$), the *T* allele attains high frequency even when it is nearly lethal in homozygous male bearers, imposing an 80% survival cost on them.

One way to resolve sexual antagonism is to restrict the expression of a trait to the sex it benefits^{13,15,19}. Counterintuitively, this is not necessarily the expected outcome for sexually antagonistic traits when they are subject to sex-linked mating preferences. For instance, the presence at high frequency of the W-linked *P* allele can select against modifiers that restrict expression of the *T* allele to females, because female-specific expression, although it increases the viability of males that bear the *T* allele, also decreases their mating success. Sex-linked preferences can thus impede the evolution of sex-specific expression and, by extension, sexual dimorphism¹⁵.

I have thus far limited the discussion to classical sexually antagonistic traits. However, the model applies more generally to three categories of costly male-specific traits: those that (i) increase the fitness of daughters; (ii) have no effect on the fitness of daughters; or (iii) act as an indicator of 'good genes' (for example, classical handicap traits). For category (i), costly male traits that increase offspring fitness are functionally identical to sexually antagonistic traits in my model. Such traits include male parental care²⁴, which is more common in ZW than XY

species²⁵. For category (ii), W-linked preferences for traits with no effect on females ($s_f = 0$) but large costs in males ($s_m \gg 0$) evolve neutrally. Such preferences can therefore drift to high frequency, which could possibly drive the evolution of exaggerated male-specific phenotypes that have previously been assumed to be the result of Fisherian runaway processes^{2,6,7}.

For category (iii), if a male-specific handicap signals intrinsic sex-independent quality^{2,6}, then a W-linked mating preference for handicapped males is favoured irrespective of the costs of the handicap, because daughters enjoy higher quality without suffering the handicap²². An analogous autosomal preference is transmitted to sons half of the time, so the higher quality of the offspring of its bearers must be offset by fitness costs in their handicapped sons. If the handicap is too costly, an autosomal mating preference for it will not spread—although a W-linked preference will. The handicap then signals a 'sexually antagonistic genome': good in females (because of the high quality it imparts) but bad in sons (because of the severe cost of the handicap). Formal modelling of this process (Supplementary Information) reveals: (i) that the W-linked preference is always favoured under the standard 'Spence condition'^{6,26} that the viability cost of the handicap is proportionally lower in higher-quality males; (ii) that more stringent conditions are required for the analogous autosomal preference to be favoured; and (iii) that the handicap must be heritable for these differences to hold.

In the above model, the selfish W-linked *P* allele can drive to high frequency a trait that severely impairs male survival. This might create selection for autosomal suppression of the preference encoded by the *P* allele. To study this possibility, I considered an augmented model with a third locus that is autosomal but is not linked to the trait locus. At this locus, there segregates a mutant allele *S* that suppresses the effect of the *P* allele, such that its female bearers are indiscriminate in mate choice (see Methods). Simulations reveal that the *S* allele invades only when the strength of the preference that it suppresses is weak, and when the trait carries a high net cost (Extended Data Figs. 3, 4). Thus, strong W-linked preferences appear to be robust to suppression.

Sex-specific chromosomes (the W or Y chromosomes) are often stereotyped as degraded and gene-poor, which would seem to diminish the possibility of their carrying preference genes. However, although the sex-specific chromosomes of therian mammals and neognath birds are indeed gene-poor, in other clades the sex-specific chromosome can vary widely in size and gene content^{14,27}. In addition, sex-specific chromosomes usually contain a non-degraded 'pseudo-autosomal region' that recombines in the heterogametic sex²⁸. Simulations reveal that preferences similar to those modelled above can fix in the pseudo-autosomal region, although only if they arise close to the border between this region and the sex-determining region (Extended Data Fig. 5).

The logic articulated above for the W chromosome applies to other genetic elements with exclusive or predominantly maternal transmission. These include mitochondria and other cytoplasmic factors^{12,13}, intracellular parasites such as *Wolbachia*²⁹ as well as microbiota, which often show vertical maternal transmission³⁰ and are known to influence behaviour—including mate choice—in a number of taxa³¹.

Although the W chromosome is sex-specific, the Z and X chromosomes are only partially sex-biased, as they are borne twice as often by one sex (males for the Z chromosome and females for the X chromosome). These transmission biases—together with recent discoveries of X- and Z-linked genes that influence mate choice (Supplementary Information)—raise the possibility that the Z and X chromosomes can shape the evolution of preferences for sexually antagonistic traits; the Z chromosome for male-beneficial, female-costly traits and the X chromosome for male-costly, female-beneficial traits. The evolution of X- and Z-linked preferences for costly male-limited traits has previously been considered²¹.

Modifying the model for X- and Z-linked preferences (Methods), I find that—in both cases—preference and trait alleles can co-evolve to high frequency (Fig. 2). This effect is stronger for the Z chromosome, despite the 'biases' of the X and Z chromosomes being symmetric.

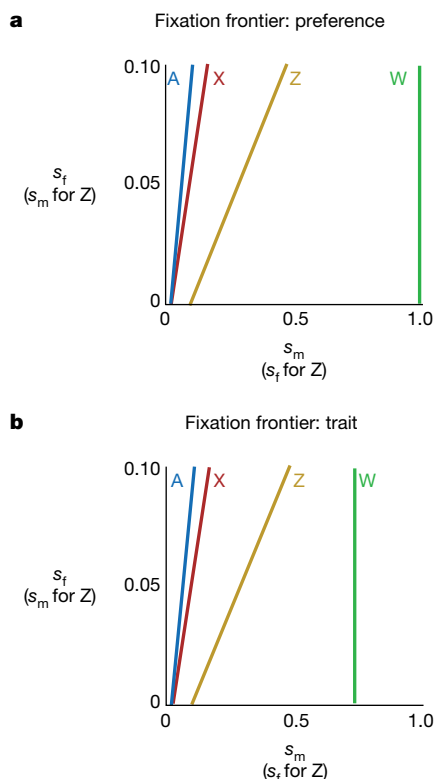


Fig. 2 | Relative propensities of the W, Z and X chromosomes to evolve female mating preferences for males that exhibit sexually antagonistic traits. a, b, The preference strength is $\alpha = 5$ in all cases. Each line is a frontier between a parameter region in which the preference (a) or trait (b) allele attains high frequency (region to the left of the line), and a parameter region in which it does not (region to the right of the line). The frontier for autosomes (labelled A) is displayed for reference. Note that Z-linked preferences are for male-beneficial, female-costly traits (contrary to the W- and X-linked preferences), so the axes are reversed for the Z chromosome. W-linked preferences for males displaying female-beneficial, male-costly traits fix for any degree of sexual antagonism. Z-linked preferences for males displaying male-beneficial, female-costly traits fix even with substantially female-costly traits, although the parameter range over which they fix is smaller than for W-linked preferences. X-linked preferences for female-beneficial, male-costly traits fix only when the degree of sexual antagonism is relatively small, although they nonetheless fix in regions in which autosomal preferences cannot. Note that Z- and X-linked preferences (unlike W-linked preferences) fix only when they also drive their preferred traits to high frequency.

To understand this, consider sex-chromosome transmission from ZW and XX females to offspring. A Z-linked allele that encodes a mating preference for a male-beneficial trait is passed on by a mother only to her sons, and thus gains an immediate advantage. By contrast, an X-linked preference allele is transmitted equally to sons and daughters, and thus immediately experiences both the cost and benefit of the trait. In fact, the pedigree transmission profiles of X- and Z-linked preference alleles, starting in females, are symmetric, except for the initial sons-only generation of the Z-linked allele (Supplementary Information), which explains why Z-linked mating preferences for sexually antagonistic traits evolve more readily. As expected, the effect is weaker for both the Z and X chromosomes than for the W chromosome (Fig. 2).

I have considered a population in which mate choice is practised exclusively by females, but the model also applies to male mate choice, which recent work has suggested is more common than has previously been recognized³².

To investigate the empirical possibility of sex-linked preferences, I collected a list of known genomic locations of mate-preference genes (Supplementary Information). Sex chromosomes are substantially over-enriched for preference genes across a variety of heterogametic

species. Sex-specific chromosomes do not feature prominently, probably because they are highly degenerate in the majority of species in the list. Indeed, one of the major goals of the theoretical work presented here is to point genomic research on mate preferences towards species with gene-rich sex-specific chromosomes.

The model described here predicts different outcomes for XY and ZW systems when mate choice is practised predominantly by females. In ZW species, the female-specific W chromosome is a very strong attractor of preferences for male-costly, female-beneficial traits, whereas the male-biased Z chromosome attracts preferences for male-beneficial, female-costly traits. By contrast, XY species have no female-specific chromosome and the X chromosome attracts preferences more weakly than does the Z chromosome (Fig. 2). Therefore, ZW species are particularly prone to the evolution of sex-linked preferences for sexually antagonistic traits. This is consistent with the phylogenetic association between ZW heterogamety and greater male ornamentation in vertebrates²³, although this relationship is ambiguous within some clades³³. Further comparative research—especially in clades with rapid heterogametic transitions—would be useful in clarifying this relationship¹⁴.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-019-1271-7>.

Received: 1 February 2018; Accepted: 8 May 2019;

Published online 5 June 2019.

1. Fisher, R. A. *The Genetical Theory of Natural Selection* (Oxford Univ. Press, 1930).
2. Zahavi, A. Mate selection—a selection for a handicap. *J. Theor. Biol.* **53**, 205–214 (1975).
3. Lande, R. Models of speciation by sexual selection on polygenic traits. *Proc. Natl Acad. Sci. USA* **78**, 3721–3725 (1981).
4. Kirkpatrick, M. Sexual selection and the evolution of female choice. *Evolution* **36**, 1–12 (1982).
5. Trivers, R. *Social Evolution* (Benjamin Cummings, 1985).
6. Grafen, A. Biological signals as handicaps. *J. Theor. Biol.* **144**, 517–546 (1990).
7. Andersson, M. B. *Sexual Selection* (Princeton Univ. Press, 1994).
8. Pomiankowski, A. N. in *Oxford Surveys in Evolutionary Biology* (eds Harvey, P. H. & Partridge, L.) 136–184 (Oxford Univ. Press, 1988).
9. Albert, A. Y. K. & Otto, S. P. Sexual selection can resolve sex-linked sexual antagonism. *Science* **310**, 119–121 (2005).
10. Kokko, H., Brooks, R., McNamara, J. M. & Houston, A. I. The sexual selection continuum. *Proc. R. Soc. Lond. B* **269**, 1331–1340 (2002).
11. Pomiankowski, A. Sexual selection: the handicap principle does work—sometimes. *Proc. R. Soc. Lond. B* **231**, 123–145 (1987).
12. Burt, A. & Trivers, R. *Genes in Conflict: The Biology of Selfish Genetic Elements* (Belknap, 2006).
13. Haig, D., Úbeda, F. & Patten, M. M. Specialists and generalists: the sexual ecology of the genome. *Cold Spring Harb. Perspect. Biol.* **6**, a017525 (2014).
14. Bachtrog, D. et al. Sex determination: why so many ways of doing it? *PLoS Biol.* **12**, e1001899 (2014).
15. Rice, W. R. Sex chromosomes and the evolution of sexual dimorphism. *Evolution* **38**, 735–742 (1984).
16. van Doorn, G. S. Intralocus sexual conflict. *Ann. NY Acad. Sci.* **1168**, 52–71 (2009).
17. Cox, R. M. & Calsbeek, R. Sexually antagonistic selection, sexual dimorphism, and the resolution of intralocus sexual conflict. *Am. Nat.* **173**, 176–187 (2009).
18. Innocenti, P. & Morrow, E. H. The sexually antagonistic genes of *Drosophila melanogaster*. *PLoS Biol.* **8**, e1000335 (2010).
19. Cheng, C. & Kirkpatrick, M. Sex-specific selection and sex-biased gene expression in humans and flies. *PLoS Genet.* **12**, e1006170 (2016).
20. Seger, J. & Trivers, R. Asymmetry in the evolution of female mating preferences. *Nature* **319**, 771–773 (1986).
21. Kirkpatrick, M. & Hall, D. W. Sexual selection and sex linkage. *Evolution* **58**, 683–691 (2004).
22. Hastings, I. M. Manifestations of sexual selection may depend on the genetic basis of sex determination. *Proc. R. Soc. Lond. B* **258**, 83–87 (1994).
23. Reeve, H. K. & Pfennig, D. W. Genetic biases for showy males: are some genetic systems especially conducive to sexual selection? *Proc. Natl Acad. Sci. USA* **100**, 1089–1094 (2003).
24. Clutton-Brock, T. H. *The Evolution of Parental Care* (Princeton Univ. Press, 1991).
25. Reeve, H. K. & Shellman-Reeve, J. S. The general protected invasion theory: sex biases in parental and alloparental care. *Evol. Ecol.* **11**, 357–370 (1997).
26. Spence, M. Job market signaling. *Q. J. Econ.* **87**, 355–374 (1973).
27. Berset-Brändli, L., Jaquière, J., Broquet, T., Ulrich, Y. & Perrin, N. Extreme heterochiasmy and nascent sex chromosomes in European tree frogs. *Proc. R. Soc. Lond. B* **275**, 1577–1585 (2008).
28. Otto, S. P. et al. About PAR: the distinct evolutionary dynamics of the pseudoautosomal region. *Trends Genet.* **27**, 358–367 (2011).

29. Werren, J. H., Baldo, L. & Clark, M. E. *Wolbachia*: master manipulators of invertebrate biology. *Nat. Rev. Microbiol.* **6**, 741–751 (2008).
30. Funkhouser, L. J. & Bordenstein, S. R. Mom knows best: the universality of maternal microbial transmission. *PLoS Biol.* **11**, e1001631 (2013).
31. Ezenwa, V. O., Gerardo, N. M., Inouye, D. W., Medina, M. & Xavier, J. B. Animal behavior and the microbiome. *Science* **338**, 198–199 (2012).
32. Edward, D. A. & Chapman, T. The evolution and significance of male mate choice. *Trends Ecol. Evol.* **26**, 647–654 (2011).
33. Mank, J. E., Hall, D. W., Kirkpatrick, M. & Avise, J. C. Sex chromosomes and male ornaments: a comparative evaluation in ray-finned fishes. *Proc. R. Soc. Lond. B* **273**, 233–236 (2006).

Acknowledgements I thank C. Veller for research assistance and comments on the manuscript. I am grateful to D. Haig, R. Trivers and J. Losos for comments on the manuscript, S. Otto, M. Nowak, M. Zuk, L. Hadany and J. Boyle for helpful discussions, and C. Noble for help with figure preparation. The simulations in this paper were run on the Odyssey cluster supported by the FAS Division of Science, Research Computing Group at Harvard University. I am supported by an NSF graduate research fellowship.

Reviewer information *Nature* thanks Andrew Pomiankowski and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Competing interests The author declares no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-019-1271-7>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-019-1271-7>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to P.M.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

METHODS

In all versions of the model considered here, the population is assumed to be infinite, with non-overlapping generations in which the order of events is: viability selection, mating, reproduction and death, followed by viability selection among the offspring, and so on. The organism is diploid with heterogametic sex determination. Mendelian segregation operates among all loci.

The mate choice model is one of fixed relative preferences^{4,20}. In general, if there are n types of male (each expressing a different degree of some trait) in proportions $p_1, p_2 \dots p_n$ at the time of mating (after viability selection), and a given female has relative preference strengths $\alpha_1, \alpha_2 \dots \alpha_n$ over the male types, then the probability that her next mate is of type i is $\alpha_i p_i / \sum_{k=1}^n \alpha_k p_k$. If this female is of type j among m female types (each expressing a different set of preferences over the male types), with female types in proportions $q_1, q_2 \dots q_m$ after viability selection, then the fraction of all mating events in the population that are between type j females and type i males is $q_j \alpha_i p_i / \sum_{k=1}^n \alpha_k p_k$.

In the case of W-linked preferences for autosomal traits, at the W-linked preference locus there segregate the wild-type p allele and the mutant P allele, while at the autosomal trait locus there segregate the wild-type t allele and mutant T allele. The T allele encodes a trait that is costly in males but beneficial in females: tt males and females have a baseline relative viability of 1; Tt males and females have viabilities of $1 - h_T s_m$ and $1 + h_T s_f$, respectively; and TT males and females have viabilities of $1 - s_m$ and $1 + s_f$, respectively. A female bearing the p allele has equal preferences over the three male genotypes, whereas a female bearing the P allele has relative preferences $1, \alpha^{h_T}$ and α over the male genotypes tt, Tt and TT , respectively. The results discussed in the main text (Figs. 1, 2) assume $h_T = 1/2$; results for $h_T = 0$ and $h_T = 1$ are given in Extended Data Fig. 1.

The justification for the specific form of the relative preference of the females bearing the P allele for Tt males (α^{h_T}) is as follows: when $h_T = 0$, such that the T allele is recessive and the trait is not expressed by Tt males, a female that bears the P allele cannot distinguish tt and Tt males—her relative preference for Tt males should therefore be 1 (α^0). When $h_T = 1$, such that the T allele is dominant, the female cannot distinguish between Tt and TT males; her relative preference for Tt males should therefore be α (α^1). Finally, in the case of exactly intermediate dominance of the T allele ($h_T = 1/2$), the preference of a female that bears the T allele for TT males over Tt males should equal the strength of her preference for Tt males over tt males; this requires that her relative preference for Tt males be $\sqrt{\alpha}$

(that is, $\alpha^{1/2}$). A similar logic will govern the choice of intermediate relative preferences in the case of Z-linked and X-linked preferences.

In the case of X-linked preferences, the viability effects of the T and t alleles in males and females are as for the case of W-linked preferences described above. The dominance of the P allele in females is denoted by h_P : pp females have equal preferences for the three male genotypes tt, Tt and TT ; Pp females have relative preferences $1, \alpha^{h_P h_T}$ and α^{h_P} ; and PP females have relative preferences $1, \alpha^{h_T}$ and α . The results discussed in the main text (Fig. 2) assume $h_T = h_P = 1/2$; the results for other possibilities are displayed in Extended Data Fig. 1.

For Z-linked preferences, the mutant T allele encodes a trait that is beneficial in males but costly in females: tt males and females have baseline relative viability 1; Tt males and females have viabilities $1 + h_T s_m$ and $1 - h_T s_f$, respectively; and TT males and females have viabilities $1 + s_m$ and $1 - s_f$, respectively. The Z-linked mutant P allele encodes a mating preference for males that bear the T allele in the same way as the W-linked preference described above.

Finally, for the case in which the preference locus is pseudo-autosomal in a ZW system, the viability effects of the T allele and preference effects of the P allele (now at a diploid locus) are as in the case of X-linked preferences, and the preference locus recombines with the sex-determining locus in a fraction r of gametes.

In the simulations, the results of which are displayed in Figs. 1, 2 and Extended Data Fig. 1, the population starts off with initial low frequencies of the mutant P and T alleles (1% each), with the loci in Hardy–Weinberg equilibrium when diploid, and in linkage equilibrium with each other. I assume that the two alleles at the trait locus mutate from one to the other at a symmetrical rate of $u = 10^{-3}$ per replication; there is no mutation at the preference locus (see Supplementary Information for a discussion of the effects of different mutation rates). From this starting configuration in each case, the population model was simulated for 5×10^6 generations (Figs. 1, 2) or 10^6 generations (Extended Data Fig. 1), and the final frequencies of the mutant P and T alleles recorded.

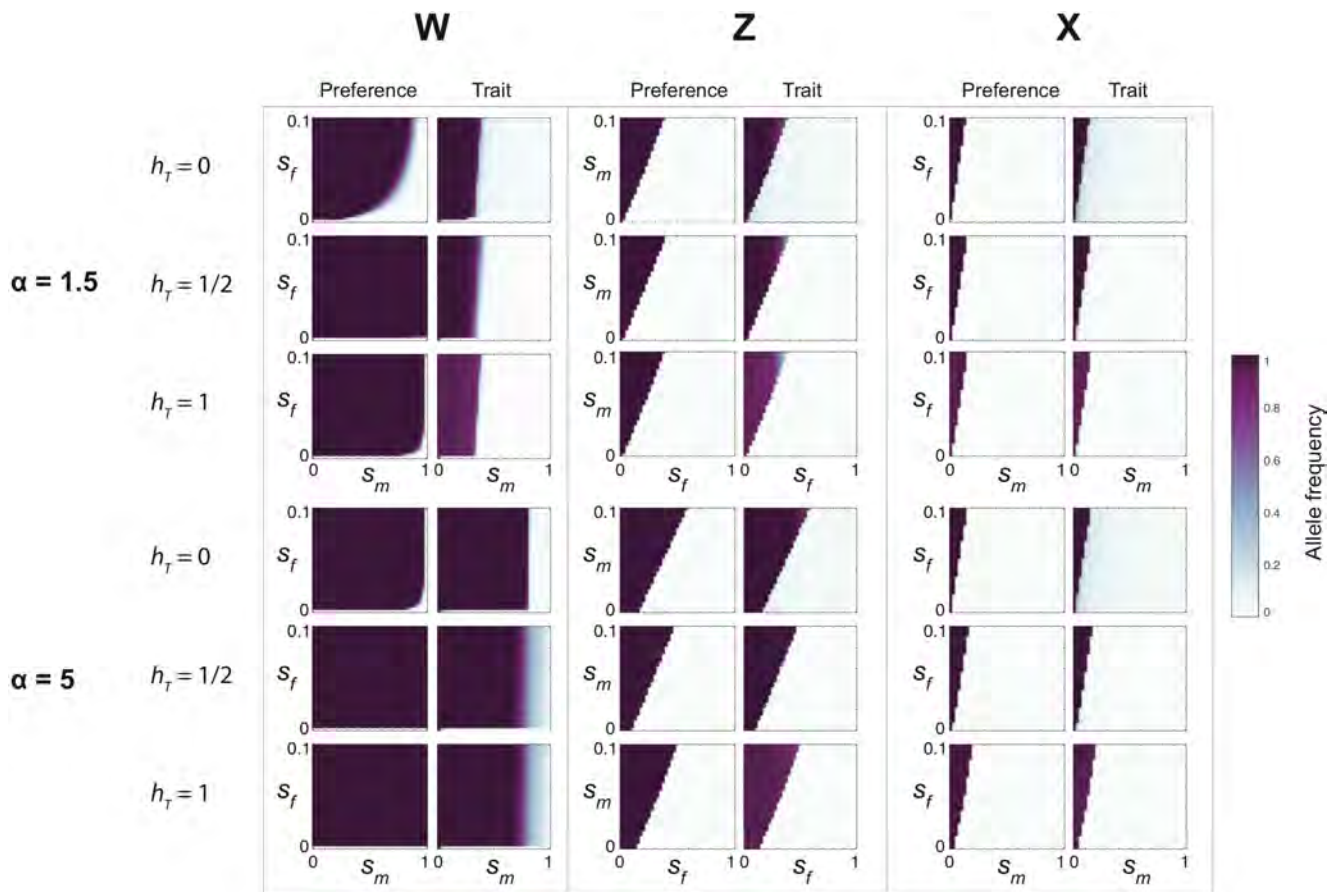
Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

No datasets were generated or analysed in this study.

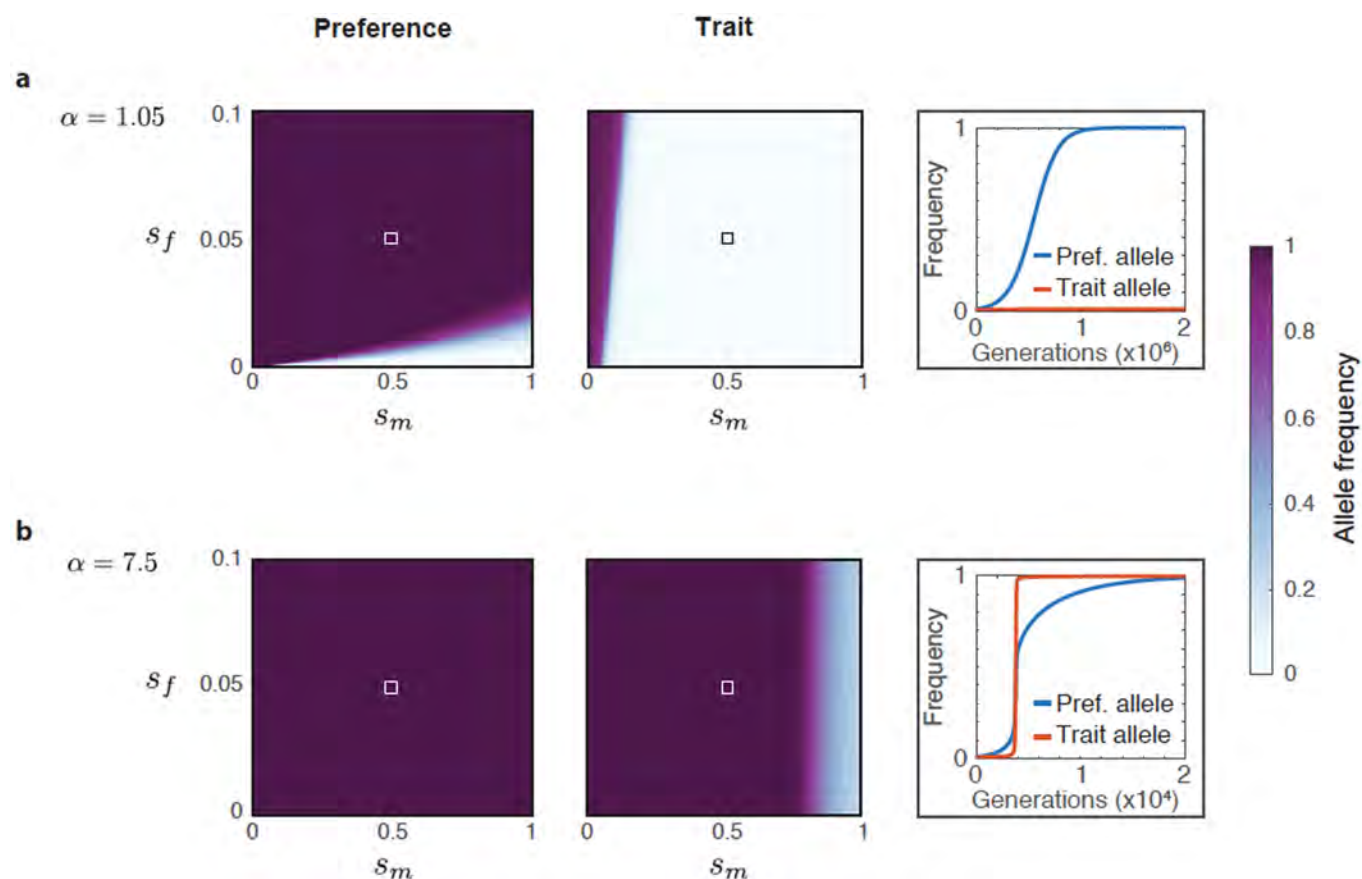
Code availability

Simulation code is available upon request.



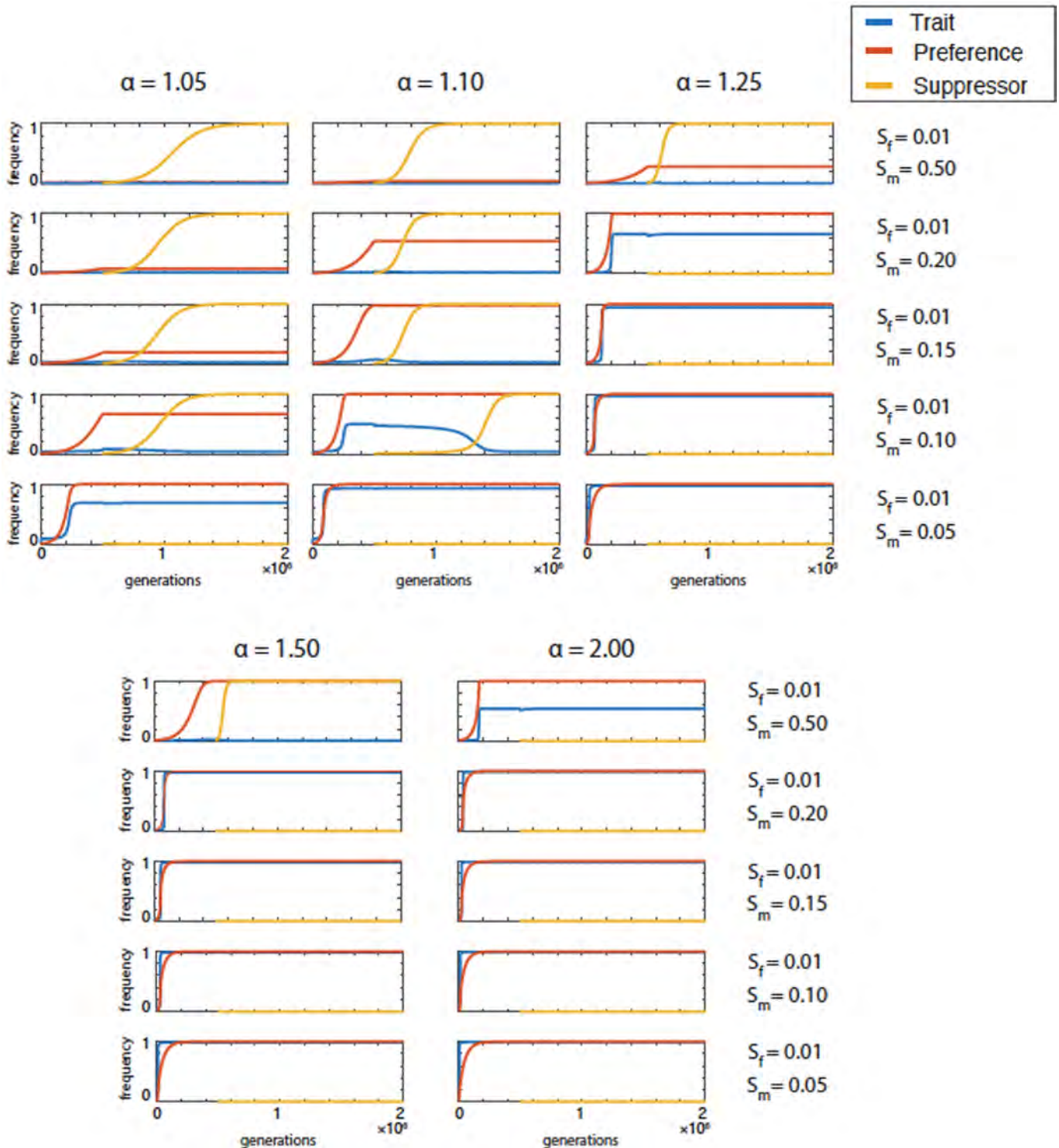
Extended Data Fig. 1 | Long-term frequencies of sex-linked preferences and traits. Frequencies of a W-linked, Z-linked and X-linked mutant P allele and an autosomal mutant T allele after 10^6 generations, each having started at 1% frequency, in Hardy-Weinberg and linkage equilibrium. The strength of the preference is $\alpha = 1.5$ (top) or $\alpha = 5$ (bottom). h_T is the dominance of the T allele with respect to the wild-type t allele; h_P is the dominance of the P allele with respect to the wild-type p allele. In the case of an X-linked preference, I assume that $h_P = h_T$; in the

case of W-linked and Z-linked preferences, h_P is not applicable, as both W- and Z-linked preferences are hemizygous in females. Note that in the case of a W-linked preference, the P allele will eventually attain high frequency in parameter regions in which it does not appear to do so here; for example, compare the results here for a W-linked preference of strength $\alpha = 1.5$ with Fig. 1c, in which frequencies after 5×10^6 generations are reported.



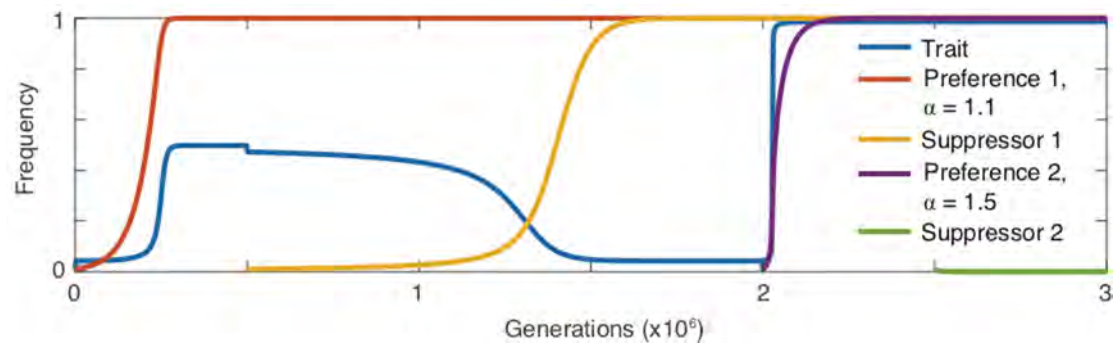
Extended Data Fig. 2 | Two equilibria for W-linked preferences for sexually antagonistic traits. Frequency trajectories of the W-linked *P* allele and an autosomal, male-costly female-beneficial *T* allele under two strengths of the preference. The heat maps displayed here are also shown in Fig. 1b (top) and Fig. 1d (bottom), and their details are described in the Methods. **a**, Sample trajectories of *P* and *T* alleles when the preference is weak and the cost of the trait to males is large. The *P* allele fixes but

the *T* allele remains at a low-frequency mutation–selection balance: the equilibrium is one in which all females prefer males that display the costly trait, but very few males display it. **b**, Sample trajectories of *P* and *T* alleles when the preference is strong. The *P* allele fixes and the *T* allele attains a very high-frequency mutation–selection balance: the equilibrium is one in which almost all males have low viability, and all females strongly prefer the low-viability males.



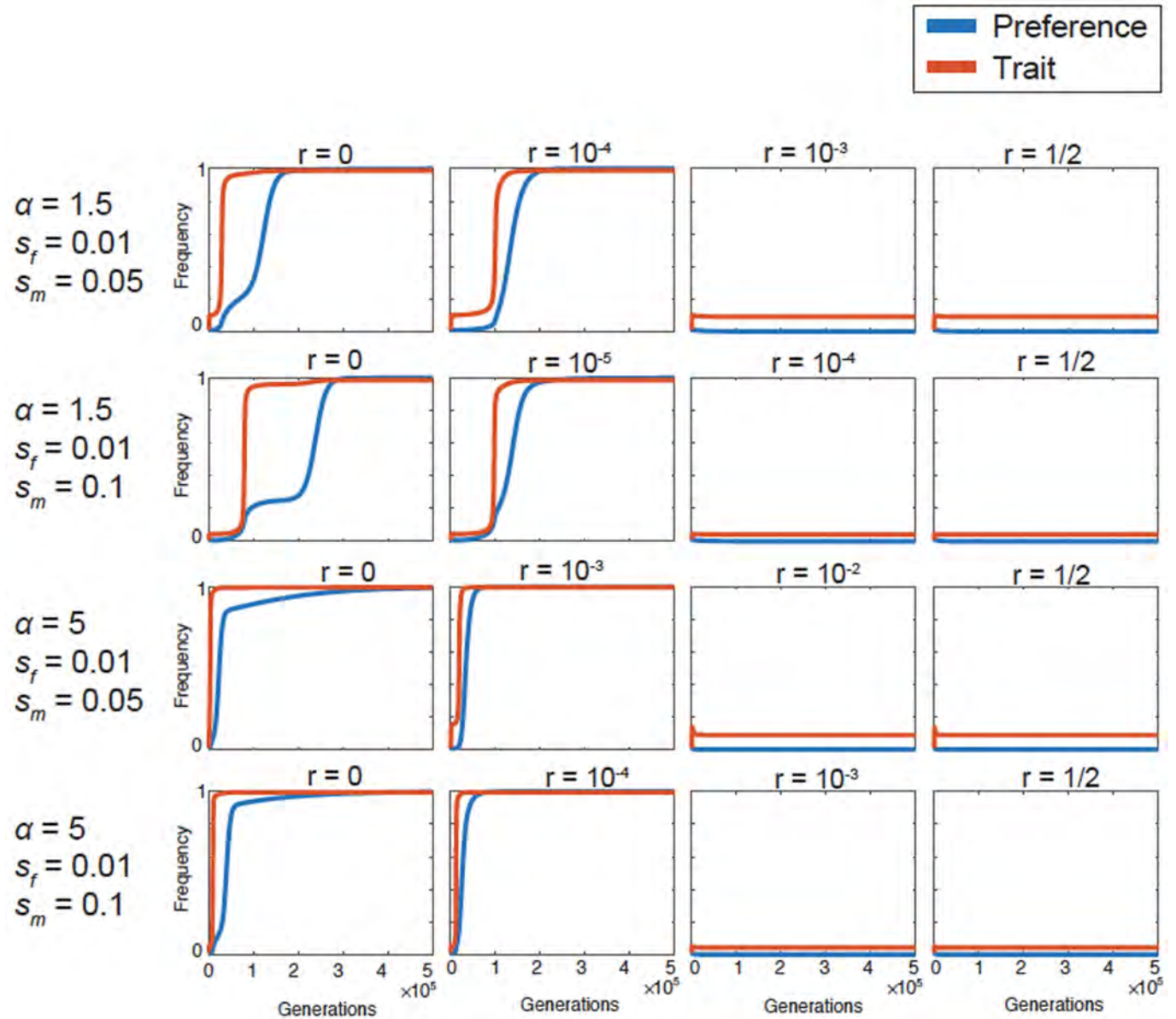
Extended Data Fig. 3 | Suppression of selfish W-linked preferences for sexually antagonistic traits. Trajectories of the W-linked P allele, the autosomal, male-costly female-beneficial T allele and an autosomal S allele that suppresses the preference allele, across various fitness effects of the trait. The mutation rate at the trait locus is 10^{-3} and the T allele is co-dominant ($h_T = 1/2$). In each simulation, at generation 0 the T and P alleles are introduced into the population. After 5×10^5 generations, a mutant S allele appears at an autosomal locus. The simulation is run for an additional 1.5×10^6 generations. Each allele is introduced at frequency

1%, in Hardy–Weinberg equilibrium if at a diploid locus, and in linkage equilibrium with respect to the other loci. The autosomal suppressor locus is unlinked to the trait locus, and the S allele is co-dominant ($h_S = 1/2$), so that a female bearing the P allele and a single S allele has preferences 1, $\alpha_i^{1/4}$ and $\alpha_i^{1/2}$ for tt , Tt and TT individuals, respectively, whereas a female with the P allele and no S allele has preferences 1, $\alpha_i^{1/2}$ and α_i for tt , Tt and TT individuals. Suppression is more likely to evolve when the strength of the W-linked preference is weak, and the average fitness cost of the trait across males and females is high.



Extended Data Fig. 4 | Arms-race dynamics between W-linked preferences and their suppressors. A weak preference allele P_1 initially invades and fixes, which pushes the sexually antagonistic T allele ($s_f = 0.01$, $s_m = 0.1$) to intermediate frequency. At 5×10^5 generations, a mutant allele that suppresses the action of P_1 appears at an unlinked locus. The suppressor invades and fixes, which eliminates the effect of

P_1 so that the T allele decreases to a low frequency. At 2×10^6 generations, a medium-strength preference allele P_2 invades and fixes, which pushes T back to a high frequency. An unlinked suppressor of P_2 appears at 2.5×10^6 generations, but immediately goes extinct: the medium-strength preference is evolutionarily resistant to suppression.



Extended Data Fig. 5 | Dynamics of preferences in the pseudo-autosomal region. Trajectories of the pseudo-autosomal mutant P allele and autosomal mutant T allele in a ZW system, for various strengths of the preference, fitness effects of the trait (always costly in males and beneficial in females) and recombination rates between the preference locus and the

sex-determining locus. The mutant trait allele is co-dominant ($h_T = 1/2$). In each case, there is some (low) threshold recombination rate, below which the preference and trait can evolve to high frequency and above which they cannot.

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- ☒ ☐ The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- ☒ ☐ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- ☒ ☐ The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- ☒ ☐ A description of all covariates tested
- ☒ ☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- ☒ ☐ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- ☒ ☐ For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
Give P values as exact values whenever suitable.
- ☒ ☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- ☒ ☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- ☒ ☐ Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection No data were collected in this paper.

Data analysis No data were analyzed in this paper.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

This paper is purely theoretical, and no data were used.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

- ☐ Life sciences ☐ Behavioural & social sciences ☒ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	This is a theoretical paper modeling the evolution of sex-linked mate preferences for sexually antagonistic traits. No data were collected or used in this paper.
Research sample	N/A
Sampling strategy	N/A
Data collection	N/A
Timing and spatial scale	N/A
Data exclusions	N/A
Reproducibility	N/A
Randomization	N/A
Blinding	N/A

Did the study involve field work? ☐ Yes ☒ No

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data

Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

Mitochondrial fragmentation drives selective removal of deleterious mtDNA in the germline

Toby Lieber^{1,3}, Swathi P. Jeedigunta², Jonathan M. Palozzi², Ruth Lehmann^{1*} & Thomas R. Hurd^{2,3*}

Mitochondria contain their own genomes that, unlike nuclear genomes, are inherited only in the maternal line. Owing to a high mutation rate and low levels of recombination of mitochondrial DNA (mtDNA), special selection mechanisms exist in the female germline to prevent the accumulation of deleterious mutations^{1–5}. However, the molecular mechanisms that underpin selection are poorly understood⁶. Here we visualize germline selection in *Drosophila* using an allele-specific fluorescent in situ-hybridization approach to distinguish wild-type from mutant mtDNA. Selection first manifests in the early stages of *Drosophila* oogenesis, triggered by reduction of the pro-fusion protein Mitofusin. This leads to the physical separation of mitochondrial genomes into different mitochondrial fragments, which prevents the mixing of genomes and their products and thereby reduces complementation. Once fragmented, mitochondria that contain mutant genomes are less able to produce ATP, which marks them for selection through a process that requires the mitophagy proteins Atg1 and BNIP3. A reduction in Atg1 or BNIP3 decreases the amount of wild-type mtDNA, which suggests a link between mitochondrial turnover and mtDNA replication. Fragmentation is not only necessary for selection in germline tissues, but is also sufficient to induce selection in somatic tissues in which selection is normally absent. We postulate that there is a generalizable mechanism for selection against deleterious mtDNA mutations, which may enable the development of strategies for the treatment of mtDNA disorders.

To visualize germline selection, we designed fluorescently labelled DNA probes that bind specifically to unique regions of the D-loops of mtDNA from either *Drosophila melanogaster* or a closely related species, *Drosophila yakuba* (Extended Data Fig. 1a–e). We then transplanted mitochondria from wild-type *D. yakuba* into a strain of *D. melanogaster* in which the mtDNA contained a temperature-sensitive point mutation in cytochrome *c* oxidase subunit I (COI^{ts})^{3,5,7}, thereby generating heteroplasmic animals that contained mixtures of wild-type and mutant mtDNA (Extended Data Figs. 1f, g, 4a, b). At the permissive temperature (18°C) the mutation does not grossly affect cytochrome oxidase activity, and is consequently not selected against in the germline^{3,5,7}. At the restrictive temperature (29°C) cytochrome oxidase activity is greatly reduced, and the mutation is selected against when paired with wild-type mtDNA from either *D. melanogaster*^{3,5,7} or *D. yakuba*⁸. This heteroplasmic animal model and mtDNA-specific fluorescent in situ-hybridization (FISH) assay enable us to directly observe and analyse mtDNA selection in vivo.

Drosophila ovaries comprise two types of tissue: germline, which gives rise to eggs and the next generation; and somatic cells, which surround the germline (Fig. 1a). Because our heteroplasmic strain contained largely mutant *D. melanogaster* mtDNA (93%), at the permissive temperature the ovaries remained largely mutant in both the germline and the soma (Fig. 1b, Extended Data Fig. 1h–h'). At the restrictive temperature, the proportion of wild-type *D. yakuba* mtDNA relative to mutant *D. melanogaster* mtDNA increased markedly in the germline but not in the soma (Fig. 1c, Extended Data Fig. 1i–i'), which demonstrates

that mtDNA selection is germline-specific. Male mtDNA is not inherited, and mtDNA FISH and quantitative PCR (qPCR) analyses of heteroplasmic *Drosophila* testes indicate that mtDNA selection is largely absent in the male germline (Fig. 1d, e, Extended Data Fig. 2). mtDNA selection is therefore female-germline-specific.

mtDNA selection is thought to occur early during oocyte development^{1–5}. In *Drosophila*, germline stem cells divide asymmetrically during this time to self-renew and to produce differentiating daughters that undergo four rounds of divisions with incomplete cytokinesis to form germline cysts (Fig. 1f). mtDNA FISH analysis showed no increase in wild-type *D. yakuba* mtDNA relative to mutant *D. melanogaster* mtDNA in germline stem cells. However, selection was observed when germ cells differentiated first into cysts and thereafter into egg chambers (Fig. 1g, i, Extended Data Fig. 3a–a''', b–b''', d–d''', e). Inhibition of cyst formation by reducing the expression of the key early differentiation factor, Bag of marbles (Bam), blocked selection (Fig. 1h, Extended Data Fig. 3c–c''', f). Our results show that mtDNA selection occurs after the stem cell stage, early in oogenesis, during germline cyst differentiation.

Germline selection could occur at the cellular level as a result of cell death. Cyst cells that inherit too many mutant mitochondrial genomes could die, and would therefore not be represented in subsequent progeny⁹. However, a previous study did not observe the death of cyst cells during selection in *Drosophila*³, and we found that inhibiting cell death by overexpressing the cell-death inhibitor p35 did not block selection (Fig. 2a). Alternatively, the unit of selection could be the mitochondrial genome. To investigate this, we tested whether expression of the *Ciona intestinalis* protein alternative oxidase (AOX)—which can partially complement loss of complex IV^{7,10}—influenced selection (Extended Data Fig. 4a–c). In effect, we bypassed the function of complex IV while leaving the mutant gene in place. Expression of AOX largely blocked selection by rescuing the mutant mitochondria (Fig. 2b, Extended Data Fig. 4d), which indicates that the selection mechanism senses defects in the oxidative phosphorylation process. Consistent with previous reports¹¹, our data show that the unit of selection is the mitochondrion itself.

We therefore asked whether morphological changes in mitochondria could be observed during selection in differentiating cysts. Using a mitochondrially targeted enhanced yellow fluorescent protein (eYFP) and live confocal microscopy, we observed that cyst mitochondria were rounder and more discrete than stem-cell mitochondria, which were more often clustered, tubular and branched (Fig. 2c, d, Supplementary Video 1). In accordance with previous findings¹², these results indicate that germline cyst mitochondria become fragmented. We propose that fragmentation enables mutant mitochondrial genomes to be distinguished from wild-type genomes. During the 2- to 8-cell cyst stage, mtDNA does not replicate³; consequently, fragmentation causes a reduction in the number of genomes per mitochondrion, which decreases the probability that both mutant and wild-type genomes reside in the same mitochondrion and improves the efficacy of selection. To facilitate selection it is also necessary for fragmentation

¹HHMI and Kimmel Center for Biology and Medicine of the Skirball Institute, Department of Cell Biology, New York University School of Medicine, New York, NY, USA. ²Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada. ³These authors contributed equally: Toby Lieber, Thomas R. Hurd. *e-mail: ruth.lehmann@med.nyu.edu; thomas.hurd@utoronto.ca

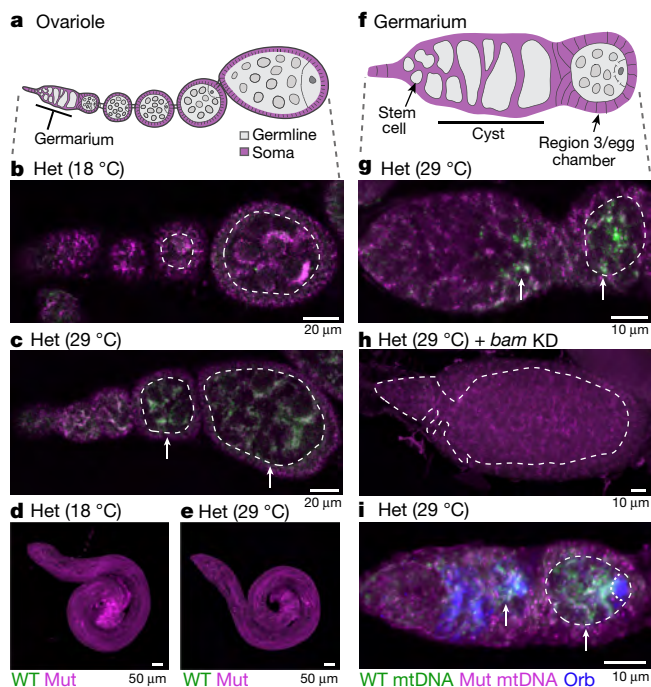


Fig. 1 | Purifying mtDNA selection is female-germline-specific and manifests during cyst differentiation. **a**, Schematic of an ovariole. The germarium is at the tip, followed by egg chambers surrounded by somatic follicle cells. **b**, **c**, Ovarioles of flies heteroplasmic (Het) for *D. melanogaster* *mt:Co1^{ts}* (Mut) and *D. yakuba* (WT) genomes hybridized with fluorescent probes that detect either wild-type or mutant genomes. Selection against the mutant genome is observed in the germline at the restrictive (29 °C, **c**) but not at the permissive (18 °C, **b**) temperature (Extended Data Fig. 1h, i). **d**, **e**, mtDNA FISH analysis of Het testes. No selection against mutant genomes is observed at either 29 °C or 18 °C (Extended Data Fig. 2). **f**, Schematic of the germarium: germline stem cells renew and produce cysts that mature into egg chambers. **g**, mtDNA FISH analysis of Het germarium at the restrictive temperature. Arrows point to wild-type mtDNA, which is first strongly detected in cyst cells (Extended Data Fig. 3a, b, **e**). **h**, mtDNA FISH analysis of Het germarium in which *bam* is knocked down (*bam* KD; resulting in reduced expression of Bam), arrested before cyst formation. No increase in wild-type mtDNA is observed (Extended Data Fig. 3c). **i**, mtDNA FISH of Het germarium co-reacted with anti-Orb antisera to mark cysts and oocytes (Extended Data Fig. 3d). **b**, **c**, **g**–**i**, The dashed lines mark the boundary between somatic and germline cells. Images here and throughout are oriented with the stem cells towards the left. For mtDNA FISH, at least eight control and experimental ovarioles, germaria or testes were imaged for each experimental condition. Imaging parameters and methodology are presented in the Methods and in Supplementary Table 3.

to prevent mitochondria from sharing their contents. To assess this we targeted a photoactivatable GFP to the mitochondrial matrix and photoactivated a subset of mitochondria in stem cells and cysts. In stem cells, the photoactivatable GFP diffused rapidly throughout the mitochondrial network, which indicates that these mitochondria share contents (Fig. 2e, g). In cysts, the photoactivatable GFP rarely passed from one mitochondrion to another (Fig. 2f, g), indicating that, at this stage, mitochondria do not readily share contents. From these observations it can be suggested that germline cyst mitochondrial fragmentation generates functionally distinguishable units for selection.

To directly test whether the fragmentation that is observed in cysts is necessary for selection, we increased the interconnectedness of cyst mitochondria by overexpressing the pro-fusion protein Mitofusin¹³ (Extended Data Fig. 5a, b). Using mtDNA FISH analysis, we found that the overexpression of Mitofusin largely abolished selection (Fig. 3a, b, Extended Data Fig. 5e–e’). Consistent with our FISH data, qPCR analysis indicated that the overexpression of Mitofusin increased the amount of mutant mtDNA while not grossly affecting the amount of wild-type

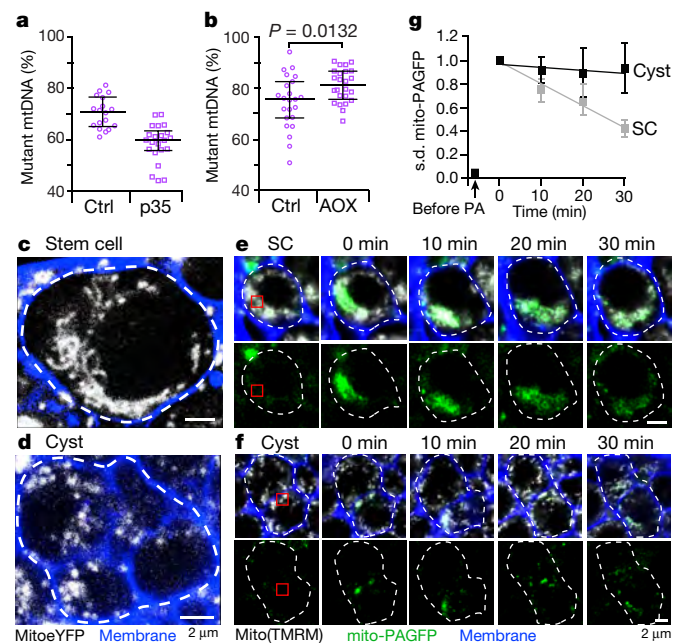


Fig. 2 | Germline cyst mitochondria undergo fragmentation.

a, **b**, Percentage of mutant mtDNA, as assayed by qPCR, in heteroplasmic ovaries with cell death blocked in the germline by expression of p35 (**a**), or with the function of the mutant complex IV bypassed by expression of the *C. intestinalis* alternative oxidase (AOX) (**b**; Extended Data Fig. 4). The mtDNA qPCR data throughout are presented as medians and interquartile range, and are compared by two-tailed unpaired *t*-tests. In Supplementary Table 2, we also present 95% confidence intervals of the difference between the control and experimental means for all datasets, and the number of biologically independent samples used to derive the statistics. **c**, **d**, Stills of live images illustrating the differing shapes of mitochondria in stem cells (**c**) and 4- to 8-cell cysts (**d**). Mitochondria, white; cell membranes, blue. Dashes outline the stem cell and 4- to 8-cell cyst (Supplementary Video 1). **e**, **f**, Time course of diffusion of the photoactivated mito-photoactivatable GFP (PAGFP) in stem cells (SC, **e**) and 4- to 8-cell cysts (**f**). Mitochondria (in the upper panels) are white; the red box marks the site of photoactivation. **g**, Quantification of the diffusion of photoactivated mito-PAGFP in stem cell and cysts. The standard deviation (s.d.) of the fluorescence intensity of GFP in the whole stem cell or the cyst at each time point was normalized to the initial post-activation value. PA, photoactivation. Data are mean \pm s.e.m. of four biological replicates.

mtDNA (Fig. 3f, Extended Data Fig. 5h). To exclude the possibility that our results were influenced by the fact that both the nuclear and mutant mitochondrial genomes were from *D. melanogaster* whereas the wild-type mitochondrial genome was from *D. yakuba*, we repeated the experiment in a heteroplasmic strain in which both wild-type and mutant mtDNAs were from *D. melanogaster*. In this *D. melanogaster*-only background, the overexpression of Mitofusin blocked selection in a similar manner (Extended Data Fig. 5i). Increasing the connectedness of cyst mitochondria by reducing expression of the pro-fission factor Drp1¹⁴ (Extended Data Fig. 5c) also blocked the selective removal of mutant mtDNA (Extended Data Fig. 5f–f’). These findings indicate that promoting mitochondrial fusion or inhibiting fission enables mutant mtDNA to hide and to escape selection during oogenesis. Therefore, a sustained fragmented phase is necessary for mtDNA selection.

To test whether mitochondrial interconnectedness could underlie the absence of mtDNA selection in germline stem cells, we promoted fragmentation in stem cells by reducing the expression of Mitofusin (Extended Data Fig. 6a, b). This induced selection in germline stem cells (Fig. 3c, Extended Data Fig. 6c–c’). A marked reduction in mutant mtDNA was observed, which suggests that—once fragmented—an elimination pathway acts to degrade mutant mtDNA. Although the reduction in expression of Mitofusin also caused defects

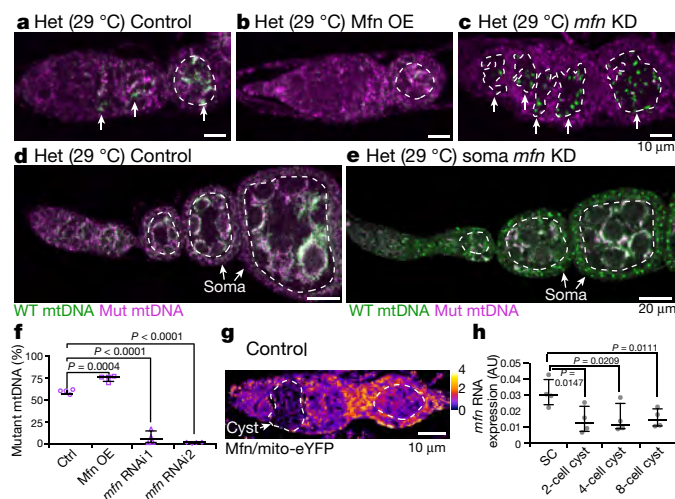


Fig. 3 | Mitochondrial fragmentation is necessary and sufficient for germline mitochondrial DNA selection. **a–c**, mtDNA FISH analysis of heteroplasmic germaria: control (**a**), germline overexpression of Mitofusin (Mfn OE; selection for wild-type mtDNA is no longer observed) (**b**) and germline knockdown of *mitofusin* (*mfn* KD, resulting in reduced expression of Mitofusin; selection for wild-type mtDNA is observed in stem cells) (**c**). Dashed circles demarcate germline. Arrows mark selection for wild-type mtDNA (Extended Data Figs. 5d, e, 6c–c’). **d, e**, The reduction of expression of Mitofusin in somatic cells is sufficient to select against mutant mtDNA in those cells. **f**, Percentage mutant mtDNA, as assayed by qPCR, of embryos laid by control heteroplasmic females and of embryos of heteroplasmic females in which Mitofusin was overexpressed (Mfn OE) or its expression was reduced in the germline (*mfn* RNAi) (Extended Data Figs. 5h, 6d, e). **g**, Germarium expressing haemagglutinin-tagged Mitofusin (Mfn), under its endogenous promoter, and mitochondrially target eYFP (mito-eYFP) reacted with anti-haemagglutinin and anti-GFP antisera. Pseudocolouring depicts the ratio of Mitofusin to mito-eYFP levels. For ratios, see the pseudocolour bar. The dashed circle indicated by the arrow outlines the cysts (Extended Data Fig. 7a–a’). **h**, *mitofusin* RNA levels decrease in cysts compared to stem cells. The RNA levels were determined by RT–qPCR and are presented as arbitrary units (AU). Data presented are median and interquartile range of four biological replicates, and were analysed using unpaired one-tailed *t*-tests (Extended Data Fig. 7b).

in germline development (Fig. 3c, Extended Data Fig. 6c–c’), qPCR analysis of mtDNA from the few young embryos obtained (Fig. 3f) and of whole ovaries (Extended Data Fig. 6d, e) confirmed that it enhanced germline selection. The overexpression of Drp1 similarly enhanced selection (Extended Data Fig. 6d, e). Control experiments indicated that Mitofusin and Drp1 do not regulate selection through processes other than fusion and fission (Extended Data Fig. 6f–h, Supplementary Note 1). Together, these data show that mitochondrial fragmentation is not only necessary but is also sufficient for germline mtDNA selection.

There is no robust selection against mutant mtDNA in most somatic tissues^{15,16}. Given that a prolonged fragmented phase is sufficient to induce selection in the germline, we asked whether it might also be sufficient in the soma. Notably, reducing Mitofusin expression induced strong selection against mutant mtDNA in somatic follicle cells (Fig. 3d, e), which demonstrates that sustained fragmentation is sufficient to induce mtDNA selection in somatic cells. Our data indicate that the key determinant, which permits selection in the germline but not the soma, is the marked decrease in fusion of germline mitochondria during early oogenesis that results in an extended phase of mitochondrial fragmentation.

To test whether this fragmented phase is caused by a decrease in Mitofusin expression, we measured the amounts of Mitofusin protein and RNA and found that both were selectively reduced in cyst mitochondria (Fig. 3g, h, Extended Data Fig. 7). Downregulation of Mitofusin was not affected by reducing the expression of known posttranslational regulators Pink1, Parkin, VCP1¹⁷ or Muli¹⁸

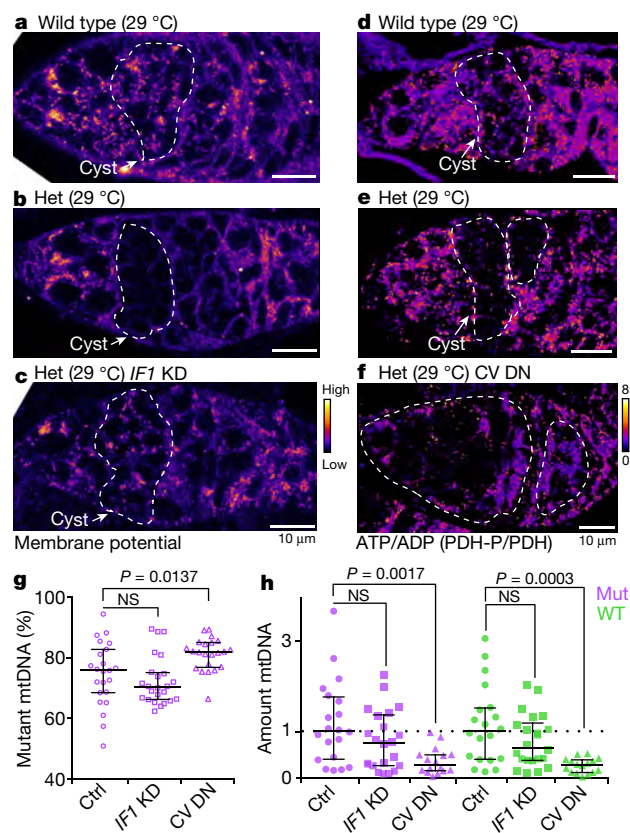


Fig. 4 | A decrease in mitochondrial ATP reduces both mutant and wild-type mtDNA. **a–c**, Pseudocoloured images of germaria of wild-type (**a**; *w¹¹¹⁸*), heteroplasmic (**b**) and heteroplasmic females with reduced expression of ATP synthase inhibitory factor 1 (CG13551, *IF1* KD) in germline (**c**) reacted with tetramethylrhodamine methyl ester to measure mitochondrial membrane potential. **d–f**, Germaria of wild-type (**d**; *w¹¹¹⁸*), heteroplasmic (**e**) and heteroplasmic females expressing a dominant negative inhibitor of complex V (CV DN) in the germline (**f**) reacted with antibodies to phosphorylated pyruvate dehydrogenase (PDH-P) and pyruvate dehydrogenase (PDH). Pseudocolouring depicts the ratio of PDH-P to PDH and is a measure ATP levels (Extended Data Fig. 9). **g, h**, Percentage mutant mtDNA (**g**) and the amount of mutant and wild-type mtDNA (**h**), as assayed by qPCR, of control heteroplasmic ovaries, of heteroplasmic ovaries in which the expression of ATP synthase inhibitory factor 1 was reduced in the germline (*IF1* KD), and of heteroplasmic ovaries in which a dominant negative inhibitor of complex V (CV DN) was expressed in the germline. In **h**, the amounts of mutant and wild-type DNA were normalized to the amounts in control ovaries.

(Extended Data Fig. 8). Mitofusin expression is therefore downregulated in germline cysts, which drives mitochondrial fragmentation and—in turn—germline mtDNA selection.

It is not known how mutant mitochondria are recognized and selected against once they are fragmented. The mitochondrial genome encodes proteins that are required for the generation of a proton motive force (PMF) and the synthesis of ATP. Mutations in mtDNA would therefore be expected to directly affect the PMF, the amount of ATP or both. Indeed, both the PMF and the amount of ATP were reduced in germline cysts, in predominantly mutant heteroplasmic flies (Fig. 4a, b, d, e). Inhibition of mitochondrial fragmentation by overexpressing Mitofusin blocked this reduction, further highlighting the importance of sustained mitochondrial fragmentation in exposing mutant genomes (Extended Data Fig. 9a–d). To determine whether a reduction in PMF marks mutant mitochondria for selection, we tested the effect of restoring the PMF in mutant mitochondria. Normally, the ATP synthase inhibitory factor 1 (IF1) prevents ATP synthase from working in reverse to restore the PMF in mutant mitochondria^{19,20}. Therefore, to restore the PMF, we reduced expression of IF1 (Fig. 4c). We observed

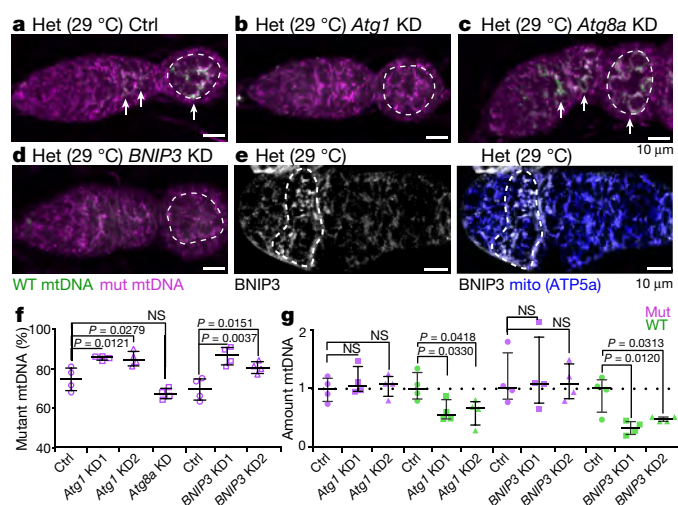


Fig. 5 | The mitophagy proteins Atg1 and BNIP3 are necessary for germline mitochondrial DNA selection. a–d, mtDNA FISH analysis of control heteroplasmic ovaries (a) and of heteroplasmic ovaries in which the expression of Atg1 (b), Atg8a (c) or BNIP3 (d) was reduced. The dashed circles demarcate the germline, and arrows point to wild-type mtDNA (Extended Data Fig. 10). e, BNIP3 protein localization in a heteroplasmic ovary; the right panel also shows mitochondria (blue) as visualized with anti-ATP5a antibody. f, Percentage of mutant mtDNA, as assayed by qPCR, of control heteroplasmic ovaries (Ctrl) and of ovaries in which the expression of Atg1 (Atg1 KD), Atg8a (Atg8a KD) or BNIP3 (CG5059 knockdown (BNIP3 KD1 and BNIP3 KD2)) was reduced. g, The amount of mutant and wild-type mtDNA, as assayed by qPCR, in ovaries in which the expression of Atg1 or BNIP3 was reduced, normalized to the amount of mutant and wild-type mtDNA in control ovaries.

no effect on selection (Fig. 4g) and no increase in the amounts of wild-type or mutant mtDNA (Fig. 4h), which indicates that a loss of PMF is not necessary for selection against mutant mitochondria. We then explored whether a reduction in the amount of mitochondrial ATP could provide a signal to select against mutant mitochondria. It is not possible to restore ATP levels in mutant heteroplasmic animals without also restoring the PMF; therefore, we tested whether reducing ATP was sufficient to make wild-type mitochondria appear mutant and promote their elimination. We generated a transgenic strain that conditionally expressed a dominant negative form of ATP synthase in both mutant and wild-type mitochondria (Fig. 4f, Extended Data Fig. 9e–g). Reduction in the amount of mitochondrial ATP reduced mutant mtDNA and, notably, wild-type mtDNA (Fig. 4g, h), which indicates that a reduction in mitochondrial ATP is sufficient to induce selection.

We next sought to determine how mutant mitochondria are selected against once they have been fragmented and their ATP has been depleted. Mitophagy would seem to be a good candidate for the mechanism, as it is the main pathway for the elimination of dysfunctional mitochondria from somatic tissues²¹. However, Parkin-mediated mitophagy has little effect on the clearance of mutant mtDNA in somatic tissues¹⁵ or in the germline⁵. Nevertheless, because mitochondrial fragmentation in stem cells caused a marked reduction in mutant mtDNA (Fig. 3c), we asked whether other mitophagy pathway components are required for germline mtDNA selection. Notably, we found that reduction in the expression of Atg1—the master regulator of autophagy²¹—blocked selection, whereas reduced expression of Atg8—a key structural component of the autophagosome that interacts with selective autophagy receptors²¹—did not (Fig. 5a–c, f, Extended Data Fig. 10a–c’). Instances of Atg1-dependent, Atg8- and Parkin-independent mitophagy have previously been described, notably in the clearance of mitochondria during the maturation of red blood cells^{22–24}, which also requires the outer mitochondrial membrane protein BNIP3L (also known as NIX)²⁵. Given these parallels, we assessed whether BNIP3 (also known as CG5059)—the *Drosophila* protein

that is most homologous to BNIP3L—was required for selection in the germline. Reducing the expression of BNIP3 inhibited selection (Fig. 5d, f, Extended Data Fig. 10d–d’). Consistent with these findings, BNIP3 is upregulated in differentiating cysts²⁶, in which it is associated with mitochondria (Fig. 5e).

Bypassing mutant complex IV (Extended Data Fig. 4d) or preventing mitochondrial fragmentation by overexpressing Mitofusin (Extended Data Fig. 5h, i) blocked selection, primarily by preventing the elimination of mutant mtDNA. However, we found that, instead of preventing the elimination of mutant mtDNA, reducing the expression of Atg1 or BNIP3 predominantly decreased the levels of wild-type mtDNA (Fig. 5g). This was also the case when Atg1 expression was reduced in a heteroplasmic strain in which both wild-type and mutant mtDNAs were from *D. melanogaster* (Extended Data Fig. 10e, f). It has previously been proposed that Pink1 inhibits the replication of mutant mtDNA, enabling wild-type mtDNA to outcompete their mutant counterparts^{3,5,11}. Our results indicate that the turnover of mitochondria is coupled to replication, such that the elimination of defective mitochondria may trigger the replication of active mitochondria and ultimately selection.

Our findings indicate that developmentally regulated fragmentation of cyst mitochondria is required to isolate their genomes and proteomes, so that mitochondria possessing mutant mtDNA can be selected against through a process requiring the mitophagy proteins Atg1 and BNIP3. Given the benefits of mitochondrial fragmentation on mtDNA selection, the question arises as to why mitochondria are not always fragmented. Enhanced fragmentation comes at a cost, as substantial reduction in the expression of Mitofusin causes mitochondrial and cellular dysfunction in both germline and somatic tissues²⁷ (Extended Data Fig. 6c–c’). In addition, evidence suggests that frequent fusion and swapping of mitochondrial content is important to maintain the health of the network²⁷ and to efficiently generate ATP²⁸. It has previously been shown that, during early oogenesis, the germline does not have a strong requirement for mitochondrially generated ATP²⁹. It may have evolved an alternative energy metabolism to tolerate the possible negative energetic consequences of reduced mitochondrial function caused by sustained fragmentation. It will be interesting to explore whether inducing mitochondrial fragmentation in somatic tissues can be used as a treatment for those suffering from mtDNA disorders. Recent work indicates that this may be the case: inducing mitochondrial fragmentation in the soma temporarily during midlife improved health and prolonged lifespan in *Drosophila* and *Caenorhabditis elegans*, possibly by promoting the removal of deleterious mtDNA^{16,30,31}. In conclusion, we have uncovered a key driver of mtDNA-purifying selection in the female germline, and our findings suggest therapeutic approaches for the treatment of mtDNA disorders.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-019-1213-5>.

Received: 4 October 2017; Accepted: 18 April 2019;

Published online 15 May 2019.

- Fan, W. et al. A mouse model of mitochondrial disease reveals germline selection against severe mtDNA mutations. *Science* **319**, 958–962 (2008).
- Stewart, J. B. et al. Strong purifying selection in transmission of mammalian mitochondrial DNA. *PLoS Biol.* **6**, e10 (2008).
- Hill, J. H., Chen, Z. & Xu, H. Selective propagation of functional mitochondrial DNA during oogenesis restricts the transmission of a deleterious mitochondrial variant. *Nat. Genet.* **46**, 389–392 (2014).
- Floros, V. I. et al. Segregation of mitochondrial DNA heteroplasmy through a developmental genetic bottleneck in human embryos. *Nat. Cell Biol.* **20**, 144–151 (2018).
- Ma, H., Xu, H. & O’Farrell, P. H. Transmission of mitochondrial mutations and action of purifying selection in *Drosophila melanogaster*. *Nat. Genet.* **46**, 393–397 (2014).
- Palozzi, J. M., Jeedigunta, S. P. & Hurd, T. R. Mitochondrial DNA purifying selection in mammals and invertebrates. *J. Mol. Biol.* **430**, 4834–4848 (2018).

7. Chen, Z. et al. Genetic mosaic analysis of a deleterious mitochondrial DNA mutation in *Drosophila* reveals novel aspects of mitochondrial regulation and function. *Mol. Biol. Cell* **26**, 674–684 (2015).
8. Ma, H. & O'Farrell, P. H. Selfish drive can trump function when animal mitochondrial genomes compete. *Nat. Genet.* **48**, 798–802 (2016).
9. Pepling, M. E. & Spradling, A. C. Mouse ovarian germ cell cysts undergo programmed breakdown to form primordial follicles. *Dev. Biol.* **234**, 339–351 (2001).
10. Fernandez-Ayala, D. J. M. et al. Expression of the *Ciona intestinalis* alternative oxidase (AOX) in *Drosophila* complements defects in mitochondrial oxidative phosphorylation. *Cell Metab.* **9**, 449–460 (2009).
11. Zhang, Y. et al. PINK1 inhibits local protein synthesis to limit transmission of deleterious mitochondrial DNA mutations. *Mol. Cell* **73**, 1127–1137.e5 (2019).
12. Cox, R. T. & Spradling, A. C. A Balbiani body and the fusome mediate mitochondrial inheritance during *Drosophila* oogenesis. *Development* **130**, 1579–1590 (2003).
13. Hwa, J. J., Hiller, M. A., Fuller, M. T. & Santel, A. Differential expression of the *Drosophila* mitofusin genes fuzzy onions (*fzo*) and *dmfn*. *Mech. Dev.* **116**, 213–216 (2002).
14. Bleazard, W. et al. The dynamin-related GTPase Dnm1 regulates mitochondrial fission in yeast. *Nat. Cell Biol.* **1**, 298–304 (1999).
15. Pickrell, A. M. et al. Endogenous Parkin preserves dopaminergic substantia nigral neurons following mitochondrial DNA mutagenic stress. *Neuron* **87**, 371–381 (2015).
16. Kandul, N. P., Zhang, T., Hay, B. A. & Guo, M. Selective removal of deletion-bearing mitochondrial DNA in heteroplasmic *Drosophila*. *Nat. Commun.* **7**, 13100 (2016).
17. Tanaka, A. et al. Proteasome and p97 mediate mitophagy and degradation of mitofusins induced by Parkin. *J. Cell Biol.* **191**, 1367–1380 (2010).
18. Yun, J. et al. MUL1 acts in parallel to the PINK1/parkin pathway in regulating mitofusin and compensates for loss of PINK1/parkin. *eLife* **3**, e01958 (2014).
19. Lefebvre, V. et al. Genome-wide RNAi screen identifies ATPase inhibitory factor 1 (ATP1F1) as essential for PARK2 recruitment and mitophagy. *Autophagy* **9**, 1770–1779 (2013).
20. Buzhynskyy, N., Sens, P., Prima, V., Sturgis, J. N. & Scheuring, S. Rows of ATP synthase dimers in native mitochondrial inner membranes. *Biophys. J.* **93**, 2870–2876 (2007).
21. Pickles, S., Vigié, P. & Youle, R. J. Mitophagy and quality control mechanisms in mitochondrial maintenance. *Curr. Biol.* **28**, R170–R185 (2018).
22. Zhang, J. et al. Mitochondrial clearance is regulated by Atg7-dependent and -independent mechanisms during reticulocyte maturation. *Blood* **114**, 157–164 (2009).
23. Zhang, J. et al. A short linear motif in BNIP3L (NIX) mediates mitochondrial clearance in reticulocytes. *Autophagy* **8**, 1325–1332 (2012).
24. Villa, E., Marchetti, S. & Ricci, J.-E. No parkin zone: mitophagy without parkin. *Trends Cell Biol.* **28**, 882–895 (2018).
25. Schweers, R. L. et al. NIX is required for programmed mitochondrial clearance during reticulocyte maturation. *Proc. Natl Acad. Sci. USA* **104**, 19500–19505 (2007).
26. Hsu, H.-J. & Drummond-Barbosa, D. A visual screen for diet-regulated proteins in the *Drosophila* ovary using GFP protein trap lines. *Gene Expr. Patterns* **23–24**, 13–21 (2017).
27. Chan, D. C. Fusion and fission: interlinked processes critical for mitochondrial health. *Annu. Rev. Genet.* **46**, 265–287 (2012).
28. Mitra, K., Wunder, C., Roysam, B., Lin, G. & Lippincott-Schwartz, J. A hyperfused mitochondrial state achieved at G1-S regulates cyclin E buildup and entry into S phase. *Proc. Natl Acad. Sci. USA* **106**, 11960–11965 (2009).
29. Teixeira, F. K. et al. ATP synthase promotes germ cell differentiation independent of oxidative phosphorylation. *Nat. Cell Biol.* **17**, 689–696 (2015).
30. Rana, A. et al. Promoting Drp1-mediated mitochondrial fission in midlife prolongs healthy lifespan of *Drosophila melanogaster*. *Nat. Commun.* **8**, 448 (2017).
31. Lin, Y.-F. et al. Maintenance and propagation of a deleterious mitochondrial genome by the mitochondrial unfolded protein response. *Nature* **533**, 416–419 (2016).

Acknowledgements We thank J. Chung, M. Guo, P. O'Farrell, H. Jacobs, H. Ma, the *Drosophila* Species Stock Center, the Bloomington *Drosophila* Stock Center and the Vienna *Drosophila* Stock Center for fly stocks; members of the Lehmann laboratory and K. Lau for discussions; Y. Abdu, L. Barton, A. Blum, S. Burden, S. Kidd, M. Murphy, A. McQuibban and D. Siekhaus for comments on the manuscript; and A. Sfeir for experimental suggestions to address the reviewers' comments. This work was supported by Canadian Institutes of Health Research grant FRN 159510 to T.R.H. and by National Institutes of Health grant R37HD41900 to R.L. T.R.H. is part of the University of Toronto Medicine by Design initiative, which receives funding from the Canada First Research Excellence Fund. R.L. is a Howard Hughes Medical Institute investigator.

Reviewer information *Nature* thanks Rachel Cox, Yukiko Yamashita and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions S.P.J. and J.M.P. contributed equally to this work. T.R.H., T.L. and R.L. designed the experiments; T.R.H., T.L., S.P.J. and J.M.P. performed the experiments; and T.R.H., T.L. and R.L. wrote the manuscript with input from all authors.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-019-1213-4>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-019-1213-4>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to R.L. or T.R.H.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

METHODS

Fly stocks. For a list of fly stocks used in this paper see Supplementary Table 1.

To generate UAS.CV-DN, the coding sequence of ATP synthase subunit C (CG1746) was amplified using Phusion High Fidelity PCR system (NEB, M0530L), the proton-accepting glutamic acid 121 mutated to a glutamine (numbered according to the start of the preprotein; E121Q) and the mutated coding sequence cloned into pVALIUM22³² using Gibson Assembly master mix (NEB, E2611S). Plasmid DNA was then injected by BestGene into a strain carrying attP40 landing sites and integrated into the second chromosome using phiC31 integrase³³.

Heteroplasmic flies. Heteroplasmic flies were generated by germ plasm transfer from either wild-type *D. yakuba* or wild-type *D. melanogaster* (*w*¹¹¹⁸) embryos into mutant *D. melanogaster* (*mt:Col^{ts} + mt:ND2^{del1}*) embryos as described previously^{5,34}. GAL4 drivers were crossed into the heteroplasmic fly lines. Heteroplasmic *yakuba/melanogaster* flies were maintained at 29 °C. Heteroplasmic *melanogaster/melanogaster* flies were maintained at 18 °C. They were mated to flies carrying RNA interference (RNAi) or overexpression constructs at 18 °C for 2–3 days. The embryos and first instar larvae were then aged for 2–3 days at 18 °C before shifting to 29 °C.

Fluorescent in situ-hybridization and immunofluorescence. *Generation of unique probes.* For *D. melanogaster*, the D-loop was amplified from mtDNA using the primers GGCCGATATCCCGCGACTGCTGGCACCATTAGTCA and GGCCGATATCCCGATCAAGGTAATCCTTTTATCAGGCA. The PCR product was digested with EcoRV and SmaI and the unique sequences were subcloned into pUC19. The DNA that was subsequently nick translated was amplified using m13 forward and reverse primers. For *D. yakuba*, the D-loop was amplified from mtDNA using the primers GGCCGATATCCCGCGACTGCTGGCACCATTGGT and GGCCAAGCTTCCCTATCAAGGTAACCTTTTATCAGGCA and subcloned into pUC19. The unique sequences were amplified using m13 forward and GATTATCTATTAATTTAGAACTTAGTATACA primers. Fluorescent probes were generated by nick translation using FISH Tag DNA kits (Thermo Fisher). The probes that recognize mtDNA of both species have previously been described³⁵.

With the exception of the temperature-shift experiments (Fig. 1b–e, Extended Data Figs. 1h, i, 2), ovaries were dissected from 1–3-day-old females. Ovaries and testes were fixed and hybridized essentially as described³⁶. Dissected ovaries and testes were fixed for 4 and 8 min, respectively, in cacodylate fixative (100 mM sodium cacodylate, pH 7.4, 100 mM sucrose, 40 mM potassium acetate, 10 mM sodium acetate, 10 mM EGTA, 5% formaldehyde). They were then washed for 4 × 10 min in 2 × SSCT (2 × SSC, 0.1% Tween-20). Ovaries were washed for 10 min in 20% then 40% formamide in 2 × SSCT and 2 × in 50% formamide in 2 × SSCT. They were then incubated in 40 µl of 2 × SSC, 50% formamide, 10% dextran sulfate, 5 µg *Escherichia coli* tRNA, 5 µg salmon sperm DNA, 40 µg BSA, and 200 ng of each fluorescent probe for 3 min at 91 °C, before incubation overnight at 30 °C. Testes were washed for 3 × 10 min in PBS/0.5% Triton X-100, dehydrated through an ethanol series, incubated in 100% ethanol overnight at 4 °C, rehydrated, incubated in 5% acetic acid at 4 °C for 5 min, washed for 3 × 5 min in PBS at 4 °C and refixed in 2% paraformaldehyde at room temperature for 55 min³⁷. After three 10 min washes in PBS/0.5% Triton X-100, testes were washed for 3 × 10 min in 2 × SSCT, exchanged into 50% formamide, and hybridized in the same way as for ovaries except that they were incubated for 4 h at 50 °C before overnight incubation at 30 °C. Both ovaries and testes were washed for 4 × 10 min in 50% formamide in 2 × SSCT at 30 °C, 10 min in 40% formamide, then 20% formamide in 2 × SSCT at room temperature, and 3 × 10 min in 2 × SSCT at room temperature. When FISH was followed by immunofluorescence, ovaries were rinsed three times in PBS and fixed again in 2% paraformaldehyde in PBS for 30 min. Immunofluorescence was carried out as previously described³⁵.

Imaging parameters are presented in Supplementary Table 3. For mtDNA FISH, at least eight control and experimental ovarioles, germaria or testes were imaged for each experimental condition. For determining the ratios of PDH-P/PDH (Fig. 4d–f), Mitofusin/mito-eYFP (Fig. 3g, Extended Data Fig. 7a) and the localization of BNIP3 (Fig. 5e, f) at least three germaria were imaged. Imaging was not done blind. Deconvolution was performed using the aggressive unsupervised profile of Huygens Professional (see Supplementary Table 3 for list of images that were deconvolved). To be able to visualize changes in germline mtDNA across samples, images were normalized so that the somatic cells of the ovary were approximately 90% mutant mtDNA, approximately the percentage determined by our qPCR measurements. Using Fiji³⁸, the display range was adjusted by modifying the minimum to remove background signal (around 10%) from the wild-type mtDNA (green) channel. Both the mutant and wild-type maximum settings were then adjusted to make the soma around 90% as measured by quantifying a region of interest in the soma after conversion to RGB colour (see Supplementary Table 3). When the soma was manipulated (Fig. 3d, e), images were instead normalized as above to make the germline approximately 65% mutant mtDNA. All the greyscale images presented in the Extended Data are non-background-subtracted, unnormalized images.

Antibodies. Primary antisera used were rabbit anti-Vasa (from the laboratory of R.L.), mouse monoclonal anti-Hts (1B1, DSHB)³⁹, mouse monoclonal anti-Orb (4H8, DSHB)⁴⁰, mouse monoclonal anti-HA (Abcam, ab130275), chicken anti-GFP (Aves Labs, GFP-1020), mouse monoclonal anti-ATP5A [15H4C4] (Abcam, ab14748), mouse anti-PDH E1 α (Abcam, ab110334) and rabbit anti-phospho-PDH E1 α (S293) (Millipore, AP1062). Secondary antibodies were DyLight 405 donkey anti-rabbit, DyLight 405 donkey anti-mouse, Cy3 donkey anti-mouse, all from Jackson ImmunoResearch, and Alexa Fluor 488 goat anti-chicken from Thermo Fisher Scientific.

qPCR quantification of mitochondrial DNA. One- to three-day-old flies were dissected. For Figs. 3f, 5f, g and Extended Data Figs. 1g, 2e, 3f, 5h, mtDNA was extracted from pools of embryos, dissected ovaries and fly carcasses as previously described⁵. Samples were mechanically homogenized with a plastic pestle in 100 µl of homogenization buffer (100 mM Tris-HCl pH 8.8, 0.5 mM EDTA, 1% SDS) and incubated for 30 min at 65 °C. Potassium acetate was added to 1 M and samples were incubated for 30 min on ice, before centrifugation at 20,000g for 15 min at 4 °C. DNA was then precipitated from the supernatant by adding 0.5 volumes of isopropanol followed by centrifugation at 20,000g for 5 min at room temperature. The resultant pellet was washed with 70% ethanol and suspended in water. qPCR was carried out using 25 ng of nucleic acid and 300 nM of each primer pair with a Roche LightCycler 480 machine and LightCycler 480 SYBR Green I Master 2X (Roche, 04887352001). The PCR program was: 10 min at 95 °C, 45 cycles of 95 °C for 15 s and 60 °C for 1 min. Dissociation curves generated through a thermal denaturation step were used to verify amplification specificity. For Figs. 2a, b, 4g, h and Extended Data Figs. 4d, 5i, 6d–f, 10e, f, individual ovaries and carcasses were homogenized in 10 mM Tris pH 8.0, 1 mM EDTA, 25 mM NaCl, 200 µg ml⁻¹ Proteinase K, incubated at 25 °C for 30 min and 95 °C for 2 min⁴¹. qPCR was carried out as described above with 1/25 of an ovary and 1/50 of a carcass. For a list of the primers used, see Supplementary Table 4.

The crossing point (Cp) values (the cycle at which the fluorescence of a sample increases above the background fluorescence) were calculated using the Second Derivative Maximum method of the Roche LightCycler 480 software. The Cp values used in the analysis were the mean values of the two primer sets that amplified the indicated genomic or mtDNAs. The amount of mutant or wild-type mtDNA = 2^{-Cp}. The percentage of mutant DNA was calculated as follows: % mutant DNA = (amount of mutant mtDNA/(amount of mutant mtDNA + amount of wild-type mtDNA)) × 100. The soma (carcass) represents the starting heteroplasmy of the animal being measured. This percentage varies from fly to fly. Because we are interested in the percentage decrease in mutant mtDNA in the germline relative to the starting heteroplasmy, the percentage mutant mtDNA in each ovary was normalized to the percentage mutant mtDNA in its corresponding carcass. The percentage wild-type mtDNA in each ovary was then derived by subtracting that value from 100%.

To determine the amount of mutant and wild-type mtDNA relative to the amount of genomic DNA, the amount of total mtDNA in each ovary was normalized to the amount of genomic DNA in that ovary. The amount of mutant and wild-type mtDNA was determined by multiplying the normalized percentage mutant or wild-type mtDNA in the ovary by the normalized amount of mtDNA in that ovary.

Because of the number of manipulations involved in generating qPCR data, any of which can result in errors, we routinely tested for outliers. Outliers were identified using the ROUT method (Q = 1%) as implemented in Prism 7 for Mac OS X GraphPad Software (<http://www.graphpad.com>). All outliers were removed ad hoc; that is, they were removed before looking at the data.

Data were analysed using unpaired two-tailed *t*-tests and 95% confidence intervals of the difference between the control and experimental means, as implemented in Prism 7 for Mac OS X GraphPad Software; see Supplementary Table 2.

Quantification of mitofusin RNA levels. To generate ovaries with 2-, 4- and 8-cell cysts, *hs-bam*; *bam*^{Δ86} flies were heat-shocked for 2 h at 37 °C in a circulating water bath, transferred to new freshly yeastified vials and incubated at 29 °C for 8 h (2-cell cyst), 22 h (4-cell cyst) and 30 h (8-cell cysts) before ovaries were extracted for total RNA isolation. Total RNA isolated from ovaries using Tri-Reagent (BioShop, TRI118) was treated with Turbo DNA-free Kit (Thermo Fisher, AM1907) to remove residual genomic DNA contamination. Reverse transcription (RT) was performed on 1 µg of total RNA using oligo(dT)20 primers (Thermo Fisher, 18418020) and Superscript III (Thermo Fisher, 18064014). qPCR was carried out using SensiFAST SYBR No-ROX qPCR kit (FroggaBio, BIO-98050) and the cycling parameters described above on 1/2 of the RT reaction with marf-specific primers. Dissociation curves generated through a thermal denaturing step were used to verify amplification specificity. Results were normalized to the mean value obtained for three genes (CG8187; CG2698; Und) with invariant expression in a range of tissues and developmental stages, as revealed by publicly available transcriptome data⁴². Data were analysed using unpaired one-tailed *t*-tests. For a list of the primers used see Supplementary Table 4.

Live imaging, photoactivation and measurement of membrane potential. For live imaging of mito-eYFP tagged mitochondria, ovaries were removed from females and the ovarioles were teased apart using tungsten needles in Halocarbon 200 oil (Halocarbon Products, 9002-83-9) on a coverslide. For photoactivation and measurement of the membrane potential, ovaries were removed from females and incubated in Schneider's medium (Life Technology, 21720) containing 1 μ M tetraphenylborate (Sigma, T25402) and 20 nM tetramethylrhodamine, methyl ester (TMRM) (Invitrogen, T668) for 30 min at room temperature in the dark. 10 μ g ml⁻¹ CellMask Deep Red plasma membrane stain (Invitrogen, C10046) was then added and the ovaries were incubated for an additional 10 min. The ovaries were washed once with Schneider's Medium before being teased apart as above in Halocarbon 200 oil on a coverslip. For all live imaging, the samples were then mounted on a slide with a gas-permeable membrane (YSI, Membrane Kit Standard) before imaging with a Zeiss LSM780 confocal microscope with Plan-Apochromat 40 \times /1.4 Oil DIC and Plan-Apochromat 63 \times /1.4 Oil DIC objectives. For all live-imaging experiments and for measurement of membrane potentials, at least three biological replicates were imaged. For photoactivation, the background signal (before photoactivation) was subtracted from all images in the time series using Fiji. Images were corrected for chromatic shifting using 0.1 nm TetraSpeck microspheres (Thermo Fisher) and deconvolved using Huygens Essential X11. For the quantification of the diffusion of photoactivated mito-PAGFP in stem cells and cysts, the standard deviation of the PAGFP fluorescence intensity was calculated using Fiji as previously described^{43,44}. An increase in diffusion of PAGFP leads to a decrease in the standard deviation and indicates an increase in the number of productive fusion events. Imaris software (Bitplane) was used to quantify mitochondrial motility in germaria. Individual mitochondria were tracked using the autoregressive motion algorithm, and for each mitochondrion the distance moved (μ m) in one second was measured (displacement delta length). Mean displacement of all mitochondria over one minute of live imaging is reported.

Clear native gel electrophoresis. The oligomerization of ATP synthase was assessed by clear native PAGE (CN-PAGE)⁴⁵. Ten pairs of ovaries were homogenized in 50 μ l PBS and mixed with 50 μ l 0.1% digitonin (Thermo Fisher, BN20061) in PBS. After incubation on ice for 15 min, samples were centrifuged (10,000g) for 15 min at 4°C. The pellets (mitoplast fraction) were washed in 200 μ l PBS and centrifuged (10,000g) for 15 min at 4°C, then solubilized in 25 μ l 1 \times NativePAGE sample buffer (Thermo Fisher, BN20032) supplemented with 10 μ l 5% digitonin (Thermo Fisher, BN20061). After incubation on ice for 15 min, samples were centrifuged (20,000g) for 30 min at 4°C. Samples (15 μ l) were then resolved on 1.0-mm, 10-well NativePAGE 3–12% Bis-Tris Gels (Thermo Fisher, BN2011BX10) with 1 \times NativePAGE running buffer (Thermo Fisher, BN2001) according to the manufacturer's instructions, at 4°C. The cathode buffer was supplemented with 0.02% (w/v) *n*-dodecyl β -D-maltoside (Sigma, D4641) and 0.05% (w/v) sodium deoxycholate (Sigma, 30970). After CN-PAGE, proteins were transferred to 0.2 μ m PVDF membranes (Bio-Rad, 162-0174) using an XCell SureLock Mini-Cell Electrophoresis System (Thermo Fisher, EI0001) in buffer comprising 48 mM Tris (Sigma, T1503), 39 mM glycine (Thermo Fisher, BP381-1), 0.05% (w/v) SDS (Sigma, L3771), 20% (v/v) methanol (Thermo Fisher, A412-4), pH 8.3. The membrane was blocked in PBS, 0.1% Tween 20 with 5% skimmed milk powder and incubated with primary antibody for 1 h at room temperature. Blots were incubated with the appropriate secondary antiserum for 1 h at room temperature, treated with Pierce ECL Western Blotting Substrate (Thermo Fisher, 32106) according to the manufacturer's instructions, and visualized on the ChemiDoc MP Imaging System (BioRad, 170-8280).

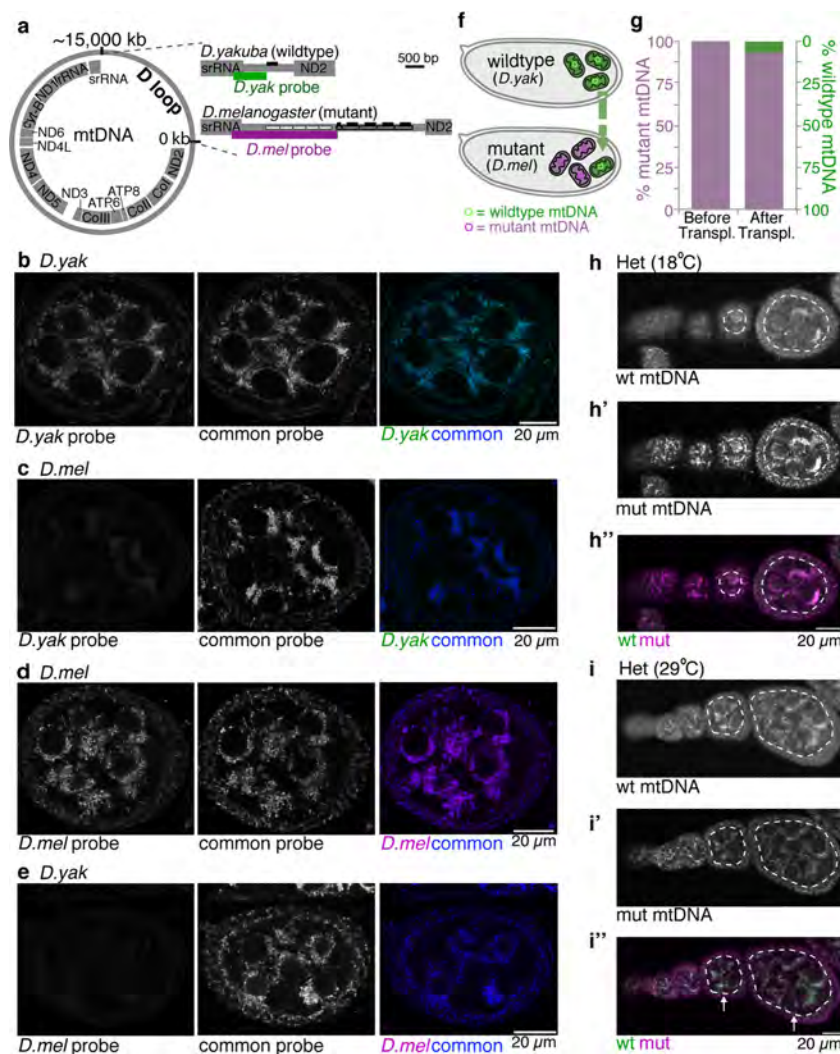
ATP/ADP determination. To measure ATP and ADP, embryos were dechorionated in bleach for 2 min, washed in PBS containing 0.1% Triton X-100, and homogenized (10 embryos per sample) with a pestle in 12 μ l Assay Buffer (Sigma, MAK135A). Samples were then analysed using ADP/ATP Ratio Assay Kit (Sigma, MAK135) according to the manufacturer's instructions. Luminescence was recorded using the Synergy H1 Microplate Reader (BioTek, BTH1M). Data were analysed using paired *t*-tests.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

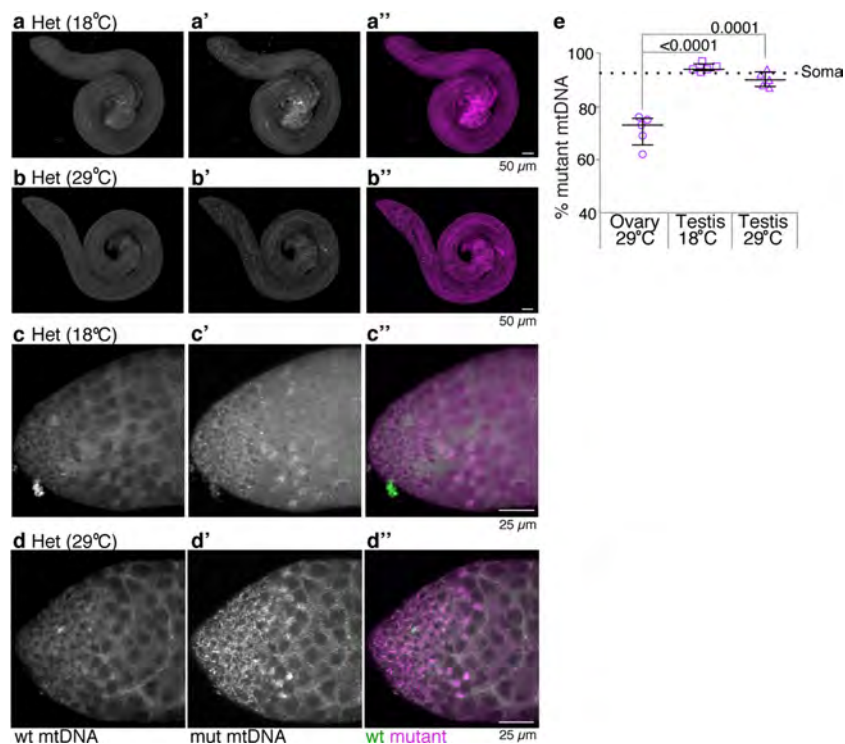
Source Data for all graphs are provided with the paper. The Cp values associated with each primer pair and DNA, and confocal image data are available upon request.

- Ni, J.-Q. et al. A genome-scale shRNA resource for transgenic RNAi in *Drosophila*. *Nat. Methods* **8**, 405–407 (2011).
- Markstein, M., Pitsouli, C., Villalta, C., Celniker, S. E. & Perrimon, N. Exploiting position effects and the gypsy retrovirus insulator to engineer precisely expressed transgenes. *Nat. Genet.* **40**, 476–483 (2008).
- Matsuura, E. T., Chigusa, S. I. & Niki, Y. Induction of mitochondrial DNA heteroplasmy by intra- and interspecific transplantation of germ plasm in *Drosophila*. *Genetics* **122**, 663–667 (1989).
- Hurd, T. R. et al. Long Oskar controls mitochondrial inheritance in *Drosophila melanogaster*. *Dev. Cell* **39**, 560–571 (2016).
- McKim, K. S., Joyce, E. F. & Jang, J. K. in *Meiosis. Methods in Molecular Biology*, Vol. 558 (ed. Keeney, S.) 197–216 (Humana Press, Totowa, 2009).
- Long, X., Colonell, J., Wong, A. M., Singer, R. H. & Lionnet, T. Quantitative mRNA imaging throughout the entire *Drosophila* brain. *Nat. Methods* **14**, 703–706 (2017).
- Schindelin, J. et al. Fiji: an open-source platform for biological-image analysis. *Nat. Methods* **9**, 676–682 (2012).
- Zaccai, M. & Lipshitz, H. D. Differential distributions of two adducin-like protein isoforms in the *Drosophila* ovary and early embryo. *Zygote* **4**, 159–166 (1996).
- Lantz, V., Chang, J. S., Horabin, J. I., Bopp, D. & Schedl, P. The *Drosophila* orb RNA-binding protein is required for the formation of the egg chamber and establishment of polarity. *Genes Dev.* **8**, 598–613 (1994).
- Gloor, G. B. et al. Type I repressors of P element mobility. *Genetics* **135**, 81–95 (1993).
- Celniker, S. E. et al. Unlocking the secrets of the genome. *Nature* **459**, 927–930 (2009).
- Gomes, L. C., Di Benedetto, G. & Scorrano, L. During autophagy mitochondria elongate, are spared from degradation and sustain cell viability. *Nat. Cell Biol.* **13**, 589–598 (2011).
- Mariotti, F. R., Corrado, M. & Campello, S. Following mitochondria dynamism: confocal analysis of the organelle morphology. *Methods Mol. Biol.* **1241**, 153–161 (2015).
- Wittig, I. & Schagger, H. Advantages and limitations of clear-native PAGE. *Proteomics* **5**, 4338–4346 (2005).
- Lewis, D. L., Farr, C. L., Farquhar, A. L. & Kaguni, L. S. Sequence, organization, and evolution of the A+T region of *Drosophila melanogaster* mitochondrial DNA. *Mol. Biol. Evol.* **11**, 523–538 (1994).
- Park, J. et al. *Drosophila* Porin/VDAC affects mitochondrial morphology. *PLoS ONE* **5**, e13151 (2010).
- Sandoval, H. et al. Mitochondrial fusion but not fission regulates larval growth and synaptic development through steroid hormone production. *eLife* **3**, e03558 (2014).
- Ni, J.-Q. et al. Vector and parameters for targeted transgenic RNA interference in *Drosophila melanogaster*. *Nat. Methods* **5**, 49–51 (2008).
- LaJeunesse, D. R. et al. Three new *Drosophila* markers of intracellular membranes. *BioTechniques* **36**, 784–790 (2004).



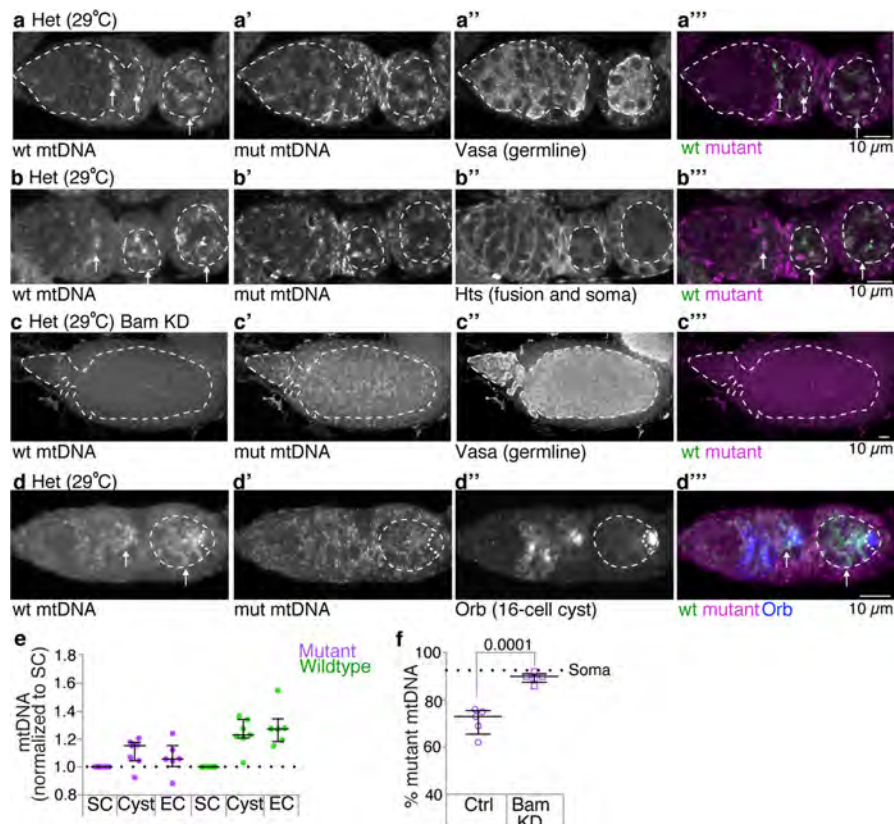
Extended Data Fig. 1 | FISH probes are specific for either *D. yakuba* or *D. melanogaster* mitochondrial DNA. **a**, Schematics of the mitochondrial genome and the D-loops of *D. yakuba* and *D. melanogaster*. In the schematic of the D-loop of *D. melanogaster*, the boxed regions denote two classes of repeated sequences. The open boxes are unique to *D. melanogaster*. The hatched boxes contain a 300-bp sequence, that is conserved in other Drosophilids⁴⁶ and is depicted by solid bars above the repeats in *D. melanogaster* and by a single solid bar above the *D. yakuba* D-loop. The FISH probes are directed against unique regions of the D-loops; the *D. yakuba*-specific probe is depicted as a green bar and the *D. melanogaster*-specific probe is depicted as a magenta bar beneath the respective D-loops. **b–e**, Confocal images of *D. yakuba* (**b, e**) and *D. melanogaster* (**c, d**) stage 7 egg chambers hybridized with *D. yakuba*-specific probes (green; **b, c**) and *D. melanogaster*-specific probes (magenta; **d, e**). All egg chambers were also hybridized with probes recognizing mtDNA of both species (common; middle panels, blue). The merged images are in the right panels. The *D. yakuba* probe hybridizes to *D. yakuba* mtDNA (**b**) but not *D. melanogaster* mtDNA (**c**). The *D. melanogaster* probe hybridizes to *D. melanogaster* mtDNA (**d**) but not *D. yakuba* mtDNA (**e**). **f**, Schematic illustrating the generation of

heteroplasmic flies by the transfer of germ plasma that contains wild-type mitochondria (green) from *D. yakuba* (*D. yak*) into *D. melanogaster* (*D. mel*) embryos that are homoplasmic for *mt:Col^{ts} + mt:ND2^{del1}* mutant mitochondria (magenta). **g**, Bar plots showing the percentage of mutant and wild-type mtDNA, as assayed by qPCR, in adult female carcasses without ovaries from the original mutant *D. melanogaster* strain and the heteroplasmic line generated by pole plasm transplantation. The data are an average of four biological replicates. **h, h', h'', i, i', i''**, Ovarioles of flies heteroplasmic (Het) for *D. melanogaster* *mt:Col^{ts}* (mut) and *D. yakuba* (wt) genomes that were shifted to 18°C (permissive temperature) for 10 days or maintained at 29°C (restrictive temperature), hybridized with fluorescent probes that detect either wild-type *D. yakuba* (green) or mutant *D. melanogaster* (magenta) genomes. Selection against the mutant genome is observed in the germline when flies were raised at 29°C. For mtDNA FISH, at least eight control and experimental ovarioles, germaria or testes were imaged for each experimental condition. Imaging parameters are presented in Supplementary Table 3. Here and in all subsequent Extended Data Figs. the greyscale images are non-background-subtracted and unnormalized.



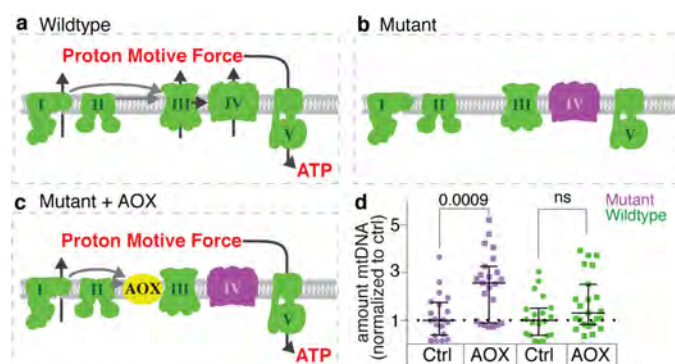
Extended Data Fig. 2 | Selection against mutant mitochondrial DNA does not occur in the male germline. a–d, Testes of heteroplasmic (Het) flies that were shifted to 18°C for 7 days (a, c) or maintained at 29°C (b, d) hybridized with fluorescent probes that detect either wild-type *D. yakuba* (green) or mutant *D. melanogaster* (magenta) genomes. The higher magnification images in c and d include the stem cells and spermatogonial cysts. Selection against mutant mtDNA is not observed in testes of flies raised at the restrictive temperature (29°C). e, Scatter plots showing the percentage of mutant mtDNA, as assayed by qPCR, of adult ovaries ($n = 5$) and testes of heteroplasmic flies raised at 29°C ($n = 5$), and

of adult testes of heteroplasmic flies shifted to 18°C for 7 days ($n = 5$). The mtDNA qPCR data throughout are presented as medians with interquartile range and compared by two-tailed unpaired *t*-tests. In Supplementary Table 2, we also present 95% confidence intervals of the difference between the control and experimental means for all datasets and the number of biologically independent samples used to derive the statistics. The dashed line denotes the percentage mutant mtDNA in whole adult-female carcasses lacking ovaries. All testes are oriented with the stem-cell niche towards the left.

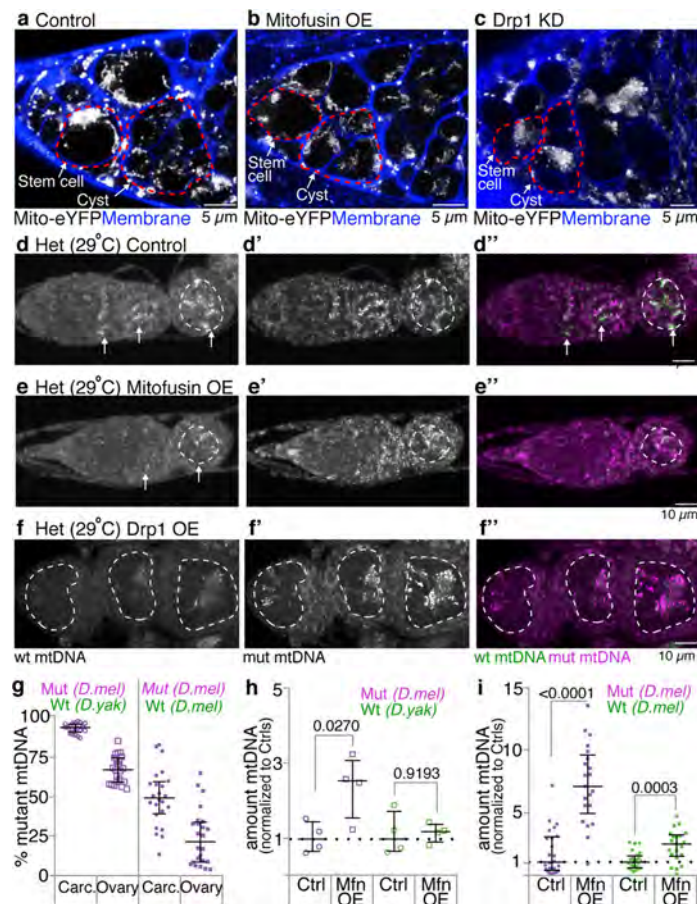


Extended Data Fig. 3 | Selection manifests in germline cyst cells and does not occur when cyst formation is blocked. **a, b**, Germaria of heteroplasmic females (Het), raised at 29°C, were hybridized with fluorescent probes that detect either wild-type *D. yakuba* or mutant *D. melanogaster* mtDNA, and reacted with anti-Vasa antisera to mark the germline (**a–a'''**) or anti-Hts (1B1) antisera to mark the fusome and somatic cells (**b–b'''**). The dashed outlines delineate the germline in the germarium (**a**), and egg chambers surrounded by somatic follicle cells (**a, b**). Wild-type mtDNA (arrows) can first be strongly detected in cysts. **c–c'''**, A germarium of a heteroplasmic fly (Het), raised at 29°C, in which cyst formation was blocked by expression of an RNAi against *bag-of-marbles* (Bam; *UAS-bam* shRNA TRiP.HMJ22155) in the germline under the control of nos-GAL4. The germarium was hybridized with fluorescent probes directed against wild-type (**c, c'''**) and mutant mtDNAs (**c', c'''**) and reacted with anti-Vasa antisera to mark the germline (**c''**). No increase in

wild-type mtDNA is observed. **d–d'''**, A germarium of a heteroplasmic fly, raised at 29°C, hybridized with fluorescent probes that detect either wild-type *D. yakuba* mtDNA (**d, d'''**) or mutant *D. melanogaster* mtDNA (**d', d'''**), and reacted with anti-Orb antisera (**d'', blue in d'''**) to demarcate all cells of the developing cysts and the oocyte in later egg chambers. Arrows in **d** and **d'''** point to wild-type mtDNA, and dashed outlines delineate the germline in the egg chambers. **e**, Scatter plots showing the relative amounts of wild-type *D. yakuba* and mutant *D. melanogaster* mtDNA, as assayed by FISH, in cysts ($n = 7$) and egg chambers (EC, $n = 6$) compared to the amount in stem cells (SC, $n = 7$). **f**, Scatter plot showing percentage of mutant mtDNA, as assayed by qPCR, of control (Ctrl; nos-GAL4 driving *UAS-mCherry* RNAi) heteroplasmic ovaries ($n = 5$) and of heteroplasmic ovaries in which cyst formation was blocked by the knockdown of *bam* (Bam KD; $n = 5$). Here and in all subsequent images, ovarioles are oriented with the stem-cell niche towards the left.

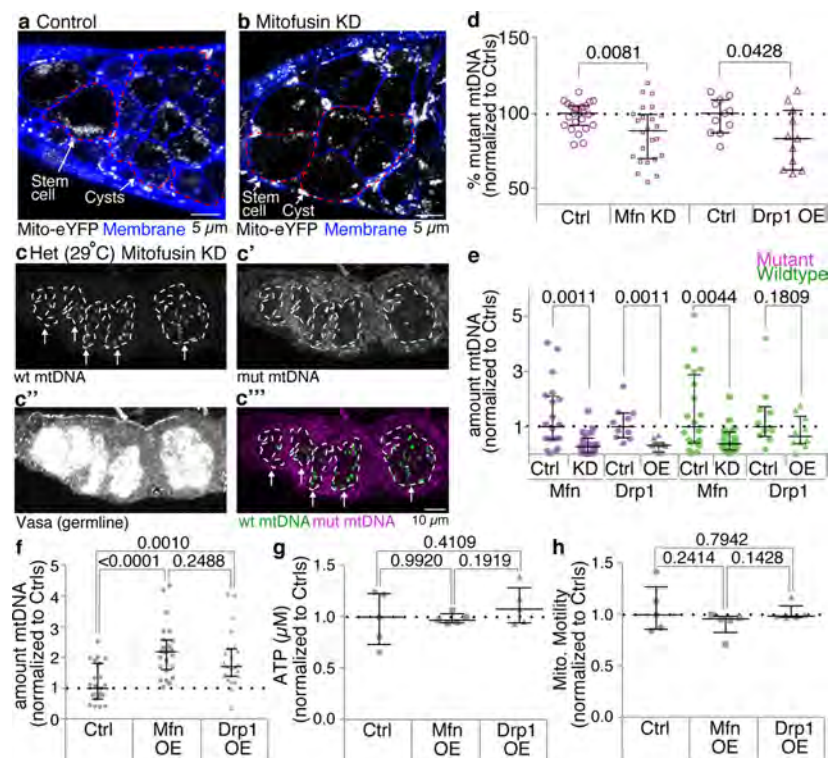


Extended Data Fig. 4 | Expression of the *C. intestinalis* alternative oxidase (AOX) rescues mutant mitochondria. **a**, In wild-type mitochondria the electron transport chain complexes (I–IV) that reside in the inner mitochondrial membrane couple the transfer of electrons to the transfer of protons across the membrane. The resulting proton motive force drives the synthesis of ATP by complex V. **b**, At the restrictive temperature the *CoI^{ts}* mutation blocks the transfer of electrons through complex IV (cytochrome oxidase, purple) resulting in the absence of both the generation of a proton motive force and ATP production. **c**, AOX (yellow) catalyses the transfer of electrons from ubiquinone to molecular oxygen, bypassing complexes III and IV. This restores the transfer of protons at complex I and the generation of ATP. **d**, Scatter plot of the amount of mutant *D. melanogaster* (purple) and wild-type *D. yakuba* (green) mtDNA, as assayed by qPCR, in ovaries expressing AOX under the control of *nos-GAL4*, normalized to the amount of mutant and wild-type mtDNA in control ovaries (Ctrl) expressing *mCherry* RNAi (Ctrl, $n = 20$; AOX, $n = 23$). Expression of AOX rescues the mutant genomes.



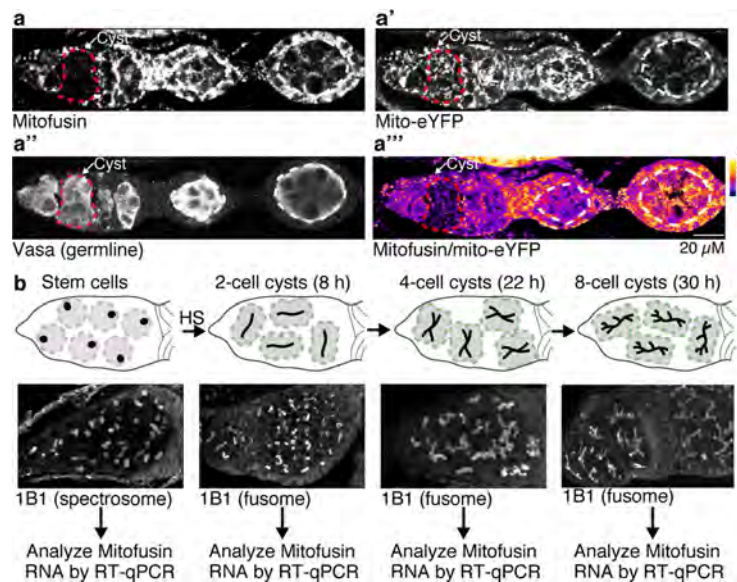
Extended Data Fig. 5 | Mitochondrial fragmentation is necessary for germline mitochondrial DNA selection. **a–c**, Stills of live images illustrating the effect that overexpressing Mitofusin (**b**) or reducing the expression of Drp1 (**c**) in the germline has on the morphology of mitochondria compared to controls (**a**, nos-GAL4 driving *UAS-mCherry* RNAi; the stills in Fig. 2c, d are higher magnifications of this image). When Mitofusin is overexpressed (nos-GAL4 driving *UAS-marf*⁴⁷), or when the expression of Drp1 is reduced (nos-GAL4 driving *UAS-Drp1*.miRNA.CDS⁴⁸), the mitochondria in the cysts are no longer discrete as they are in control cysts. The mitochondria (white) were labelled with a mitochondrially targeted eYFP and cell membranes (blue) were labelled with CellMask Deep Red Plasma membrane Stain. Stem cells and cysts are outlined in red. **d, d', d''**, Germarium of a control heteroplasmic female (nos-GAL4 driving *UAS-mCherry* RNAi), raised at 29°C, hybridized with fluorescent probes that detect either wild-type *D. yakuba* mtDNA (greyscale in **d**; green in **d''**) or mutant *D. melanogaster* mtDNA (greyscale in **d'**; magenta in **d''**). Selection for wild-type mtDNA is observed as indicated by the arrows in **d** and **d''**. **e, e', e'', f, f', f''**, Selection for wild-type mtDNA is no longer observed when Mitofusin (Mfn) is overexpressed

(nos-GAL4 driving *UAS-marf*) or when the expression of Drp1 is reduced (nos-GAL4 driving *UAS-Drp1*.miRNA.CDS). Wild-type *D. yakuba* mtDNA, greyscale in **e, f**, green in **e'', f''**; mutant *D. melanogaster* mtDNA, greyscale in **e', f'**, magenta in **e'', f''**. The dashed outlines delineate the germline. **g**, Scatter plot showing the percentage of mutant *D. melanogaster* mtDNA, as assayed by qPCR, in carcasses (carc.) and ovaries of heteroplasmic flies in which the wild-type mtDNA was either from *D. yakuba* or *D. melanogaster*. *mCherry* RNAi was expressed in the ovaries under control of nos-GAL4. **h**, Scatter plot of the amount of mutant *D. melanogaster* (purple) and wild-type *D. yakuba* (green) mtDNA, as assayed by qPCR, of young embryos laid by heteroplasmic females in which Mitofusin was overexpressed in the germline (Mfn OE, $n = 4$) normalized to the amount of mutant and wild-type mtDNA in young embryos laid by control heteroplasmic females (Ctrl; nos-GAL4 driving *UAS-mCherry* RNAi, $n = 4$). **i**, Same as **g**, except the analysis was performed on ovaries in which both wild-type and mutant mtDNAs were from *D. melanogaster* (Ctrl, $n = 24$; Mfn OE, $n = 21$). Mitofusin overexpression increases the levels of mutant mtDNA.



Extended Data Fig. 6 | Mitochondrial fragmentation is sufficient for germline mitochondrial DNA selection. **a, b**, Stills of live images illustrating the effect that reducing the expression of Mitofusin in the germline (**b**) has on the morphology of mitochondria compared to controls (**a**). When *mitofusin* is knocked down (nos-GAL4 driving *UAS-mfn* shRNA2 TRiP.HMC03883³²) the mitochondria in the stem cells are fragmented. The mitochondria (white) were labelled with a mitochondrially targeted eYFP and cell membranes (blue) were labelled with CellMask Deep Red Plasma membrane Stain. Stem cells and cysts are outlined in red. **c, c', c'', c'''**, The knockdown of *mitofusin* in the germline by expressing *mitofusin* RNAi (nos-GAL4 driving *UAS-mfn* shRNA2 TRiP.HMC03883) results in selection for wild-type mtDNA (green) occurring in stem cells. The germarium was also reacted with anti-Vasa antiserum (**c''**) to mark the germline and delineate the stem cells and cysts. Wild-type *D. yakuba* mtDNA, greyscale in **c**, green in **c'''**; mutant *D. melanogaster* mtDNA, greyscale in **c'**, magenta in **c'''**. Mutant mtDNA is readily detected in the soma but not in the germline. **d**, Scatter plot comparing the percentage of mutant mtDNA, as assayed by qPCR, of ovaries in which *mitofusin* was weakly knocked down in the germline (Mfn KD; nos-GAL4 driving *UAS-mfn* long hairpin RNA1 TRiP.JF01650⁴⁹; Ctrl, $n = 23$; Mfn KD, $n = 24$) and of ovaries in which Drp1 was overexpressed in the germline (Drp1 OE; nos-GAL4 driving *UAS-Drp1.E*; Ctrl, $n = 11$; Drp1 OE, $n = 11$). The percentage of mutant mtDNA in each case was normalized to the percentage of mutant mtDNA in control ovaries to

illustrate that the overexpression of Drp1 enhances selection to a similar extent as does a weak reduction in the expression of Mitofusin. **e**, Scatter plot of the amount of mutant *D. melanogaster* (purple) and wild-type *D. yakuba* (green) mtDNA, as assayed by qPCR, in ovaries in which the expression of Mitofusin was weakly reduced (Ctrl, $n = 20$; Mfn OE, $n = 21$) or in which Drp1 was overexpressed in the germline (Ctrl, $n = 10$; Drp1 OE, $n = 10$), normalized to the amount of mutant and wild-type mtDNA in control ovaries (Ctrl) expressing *mCherry* RNAi in the germline. Reducing the expression of Mitofusin or overexpressing Drp1 results in a decrease in mutant mtDNA. **f–h**, The effect of germline overexpression of Mitofusin (Mfn) and Drp1 on copy number (**f**), ATP levels (**g**), and mitochondrial motility (**h**) in homoplasmic wild-type *D. melanogaster* ovaries (see Supplementary Note 1). **f**, Scatter plot of the amount of mtDNA, as assayed by qPCR, in homoplasmic ovaries in which Mfn ($n = 24$) or Drp1 ($n = 24$) was overexpressed in the germline, normalized to the amount of mtDNA in control ovaries (Ctrl; nos-GAL4 driving *UAS-mCherry* RNAi, $n = 23$). **g**, Scatter plot of the amount of ATP in homoplasmic ovaries overexpressing Mfn ($n = 5$) or Drp1 ($n = 5$) in the germline under control of Maternal α -Tubulin Gal4, normalized to the amount of ATP in control ovaries (Ctrl; Maternal α -Tubulin Gal4 driving *UAS-mCherry* RNAi, $n = 5$). **h**, Scatter plot of mitochondrial motility in homoplasmic ovaries overexpressing Mfn ($n = 5$) or Drp1 ($n = 5$) in the germline. Motility was assessed by measuring mean mitochondrial displacement using live confocal microscopy and Imaris analysis software.

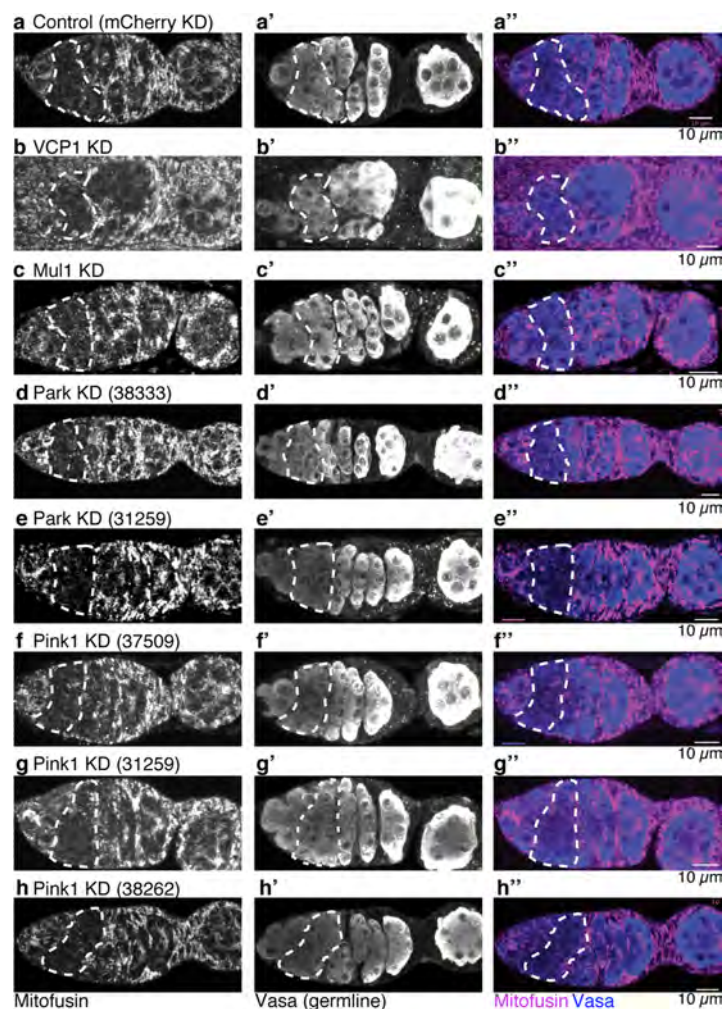


Extended Data Fig. 7 | Mitofusin is downregulated in germline cysts.

a, a', a'', a''', A germarium of a female fly expressing haemagglutinin-tagged Mitofusin (Mfn), under control of the Mitofusin promoter (Marf-gHA⁴⁸), and mitochondrially targeted eYFP (mito-eYFP⁵⁰), was reacted with anti-haemagglutinin antisera to detect Mitofusin (**a**), anti-GFP antisera to detect mitochondria (**a'**) and anti-Vasa antisera to delineate the germline (**a''**). In **a'''**, the ratio of the levels of Mitofusin to mito-eYFP is presented in pseudocolour. The colours correspond to the ratios indicated on the pseudocolour bar. The dashed red circles outline the cysts and the dashed white circles demarcate the germline in the egg chambers.

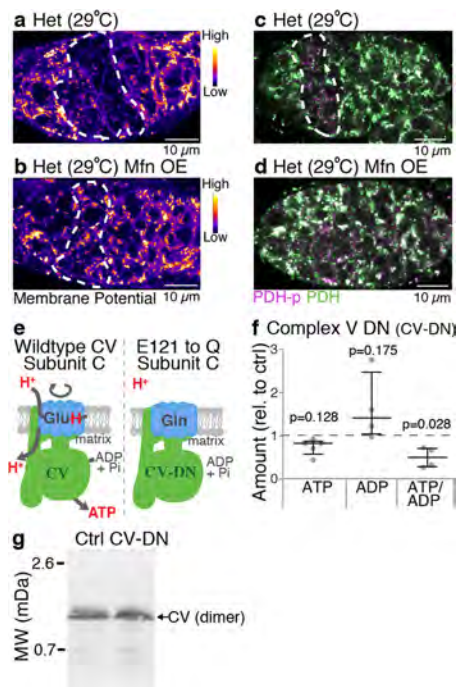
b, Scheme for quantifying the levels of *mitofusin* RNA at different time

points during early oogenesis. Females mutant for the differentiation factor Bam—which is required for cyst formation—and carrying a rescuing transgene expressing Bam under the control of a heat-shock promoter were heat-shocked at 37 °C for 2 h, and then allowed to recover for the indicated times. This enables the isolation of ovaries that contain staged cysts, predominantly at the 2-, 4- or 8-cell cyst stage. The morphology of the spectrosome and fusome, as revealed by staining with anti-Hts (1B1) antisera, was used to confirm the staging. RNA for RT-qPCR was isolated from ovaries from flies before heat shock and at the indicated times following heat shock.

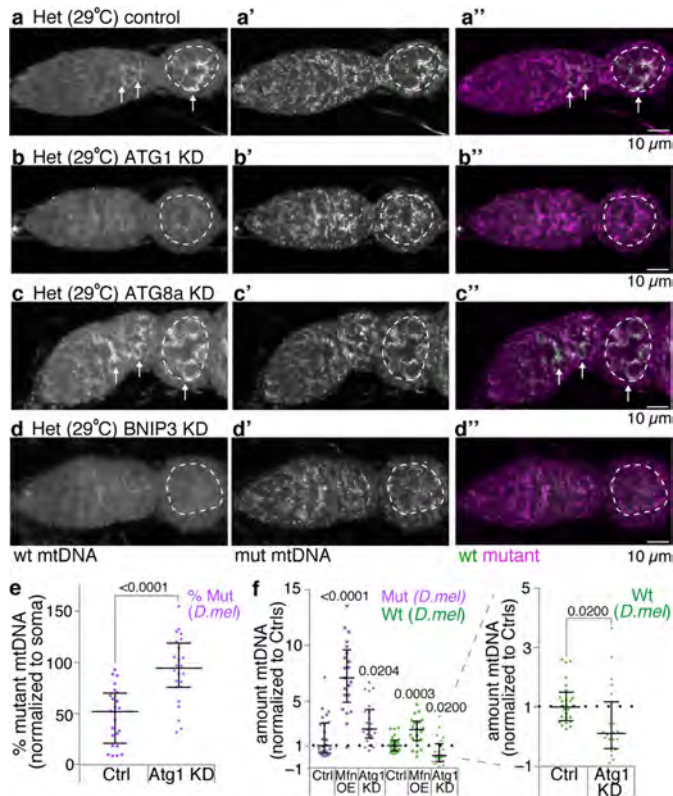


Extended Data Fig. 8 | The downregulation of Mitofusin in cysts is not mediated by known regulators of Mitofusin protein. **a–h''**, Germaria of females expressing haemagglutinin-tagged Mitofusin, under control of the Mitofusin promoter (Marf-gHA⁴⁸) reacted with anti-haemagglutinin antisera to detect Mitofusin (greyscale in **a–h**, magenta in **a''–h''**) and anti-Vasa antibody to delineate the germline (greyscale in **a'–h'**, blue in **a''–h''**).

The indicated known regulators of Mitofusin protein levels were knocked down in the germline using RNAi under the control of nos-GAL4. The numbers in parentheses are Bloomington *Drosophila* Stock Center stock numbers. All ovarioles are oriented with the stem-cell niche towards the left.



Extended Data Fig. 9 | Inhibiting mitochondrial fragmentation blocks the decrease in proton motive force and ATP levels in cysts of heteroplasmic flies. **a–d**, Germaria of heteroplasmic control flies (**a, c**; w^{1118}) and heteroplasmic flies in which Mitofusin was overexpressed in the germline (**b, d**; nos-GAL4 driving $UAS-marf^{A7}$), reacted with TMRM to visualize mitochondrial membrane potential (pseudocoloured in **a, b**) or with antibodies to phosphorylated pyruvate dehydrogenase (PDH P, purple) and pyruvate dehydrogenase (PDH, green) to measure ATP levels (**c, d**). **e**, Diagram showing the essential glutamate at position 121 in c-ring subunit that acts as the proton donor and acceptor in the proton translocation pathway. In the dominant negative c-ring (CV-DN) this glutamate was mutated to a glutamine, which can no longer bind the protons. **f**, Scatter plot illustrating the reduction in ATP/ADP ratio in embryos laid by mothers expressing CV-DN in the germline under the control of Maternal α -Tubulin Gal4. The ratios were measured using an ADP/ATP Ratio Assay Kit (Abcam ab65313). Data presented are median and interquartile range and were analysed using paired t -tests (ATP, $n = 5$; ADP, $n = 4$; ATP/ADP, $n = 4$). **g**, Blue native polyacrylamide gel illustrating that the expression of the dominant negative inhibitor of complex V (CV-DN) does not disrupt the Complex V dimer. For gel source data, see Supplementary Fig. 1.



Extended Data Fig. 10 | The mitophagy proteins Atg1 and BNIP3 are necessary for germline mitochondrial DNA selection. **a–d''**, Germaria of a control heteroplasmic female (**a, a''**; nos-GAL4 driving *UAS-mCherry* RNAi) and of heteroplasmic females in which the expression of Atg1 (**b, b''**), Atg8a (**c, c''**) or BNIP3 (**d, d''**) was reduced in the germline, raised at 29°C, hybridized with fluorescent probes that detect either wild-type *D. yakuba* mtDNA (greyscale in **a–d**, green in **a''–d''**) or mutant *D. melanogaster* mtDNA (greyscale in **a'–d'**, magenta in **a''–d''**). The dashed circles demarcate the germline in the early egg chambers. The arrows point to wild-type mtDNA. **e, f**, Scatter plots showing the percentage of mutant mtDNA and the amount of mutant (magenta) and wild-type (green) mtDNA, as assayed by qPCR, of control heteroplasmic ovaries (Ctrl, $n = 24$ in **e, f**) and of ovaries in which the expression of Atg1 was reduced in the germline (*Atg1* KD, $n = 24$ in **e**; $n = 23$ in **f**). In the left panel in **f**, the amount of mutant and wild-type mtDNA of heteroplasmic ovaries overexpressing Mitofusin (Mfn OE) is plotted to illustrate that overexpressing Mitofusin primarily inhibits selection by increasing the amount of mutant mtDNA, whereas reduced expression of Atg1 primarily inhibits selection by decreasing the amount of wild-type mtDNA. The control and Mitofusin-overexpression data are the same as that presented in Extended Data Fig. 5h. All the dissections and analyses were carried out at the same time. The right panel of **f** is a magnified view to illustrate the effect of reducing the expression of Atg1 on the level of wild-type mtDNA. In **e** and **f**, both wild-type and mutant mtDNAs were from *D. melanogaster*.

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a Confirmed

- ☐ ☒ The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- ☐ ☒ An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- ☐ ☒ The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- ☒ ☐ A description of all covariates tested
- ☒ ☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- ☒ ☐ A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- ☒ ☐ For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
Give P values as exact values whenever suitable.
- ☒ ☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- ☒ ☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- ☒ ☐ Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated
- ☐ ☒ Clearly defined error bars
State explicitly what error bars represent (e.g. SD, SE, CI)

Our web collection on [statistics for biologists](#) may be useful.

Software and code

Policy information about [availability of computer code](#)

Data collection

NA

Data analysis

NA

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

NA

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/authors/policies/ReportingSummary-flat.pdf](https://www.nature.com/authors/policies/ReportingSummary-flat.pdf)

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	in each figure or supplemental table 2
Data exclusions	explained in methods
Replication	in each figure or supplemental table 2
Randomization	not applicable
Blinding	no blinding

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
<input type="checkbox"/>	<input type="checkbox"/> Unique biological materials
<input type="checkbox"/>	<input checked="" type="checkbox"/> Antibodies
<input type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input type="checkbox"/>	<input type="checkbox"/> Palaeontology
<input type="checkbox"/>	<input checked="" type="checkbox"/> Animals and other organisms
<input type="checkbox"/>	<input type="checkbox"/> Human research participants

Methods

n/a	Involved in the study
<input type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

Unique biological materials

Policy information about [availability of materials](#)

Obtaining unique materials	in material and method section
----------------------------	--------------------------------

Antibodies

Antibodies used	in materials an method section
Validation	Describe the validation of each primary antibody for the species and application, noting any validation statements on the manufacturer's website, relevant citations, antibody profiles in online databases, or data provided in the manuscript.

Eukaryotic cell lines

Policy information about [cell lines](#)

Cell line source(s)	non used
Authentication	Describe the authentication procedures for each cell line used OR declare that none of the cell lines used were authenticated.
Mycoplasma contamination	Confirm that all cell lines tested negative for mycoplasma contamination OR describe the results of the testing for mycoplasma contamination OR declare that the cell lines were not tested for mycoplasma contamination.

Commonly misidentified lines
(See [ICLAC](#) register)

Name any commonly misidentified cell lines used in the study and provide a rationale for their use.

Palaeontology

Specimen provenance

Provide provenance information for specimens and describe permits that were obtained for the work (including the name of the issuing authority, the date of issue, and any identifying information).

Specimen deposition

Indicate where the specimens have been deposited to permit free access by other researchers.

Dating methods

If new dates are provided, describe how they were obtained (e.g. collection, storage, sample pretreatment and measurement), where they were obtained (i.e. lab name), the calibration program and the protocol for quality assurance OR state that no new dates are provided.

☐ Tick this box to confirm that the raw and calibrated dates are available in the paper or in Supplementary Information.

Animals and other organisms

Policy information about [studies involving animals](#); [ARRIVE guidelines](#) recommended for reporting animal research

Laboratory animals

in materials and methods

Wild animals

Provide details on animals observed in or captured in the field; report species, sex and age where possible. Describe how animals were caught and transported and what happened to captive animals after the study (if killed, explain why and describe method; if released, say where and when) OR state that the study did not involve wild animals.

Field-collected samples

For laboratory work with field-collected samples, describe all relevant parameters such as housing, maintenance, temperature, photoperiod and end-of-experiment protocol OR state that the study did not involve samples collected from the field.

Human research participants

Policy information about [studies involving human research participants](#)

Population characteristics

Describe the covariate-relevant population characteristics of the human research participants (e.g. age, gender, genotypic information, past and current diagnosis and treatment categories). If you filled out the behavioural & social sciences study design questions and have nothing to add here, write "See above."

Recruitment

Describe how participants were recruited. Outline any potential self-selection bias or other biases that may be present and how these are likely to impact results.

ChIP-seq

Data deposition

☐ Confirm that both raw and final processed data have been deposited in a public database such as [GEO](#).

☐ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links

May remain private before publication.

For "Initial submission" or "Revised version" documents, provide reviewer access links. For your "Final submission" document, provide a link to the deposited data.

Files in database submission

Provide a list of all files available in the database submission.

Genome browser session

(e.g. [UCSC](#))

Provide a link to an anonymized genome browser session for "Initial submission" and "Revised version" documents only, to enable peer review. Write "no longer applicable" for "Final submission" documents.

Methodology

Replicates

Describe the experimental replicates, specifying number, type and replicate agreement.

Sequencing depth

Describe the sequencing depth for each experiment, providing the total number of reads, uniquely mapped reads, length of reads and whether they were paired- or single-end.

Antibodies

Describe the antibodies used for the ChIP-seq experiments; as applicable, provide supplier name, catalog number, clone name, and lot number.

Peak calling parameters

Specify the command line program and parameters used for read mapping and peak calling, including the ChIP, control and index files used.

Data quality

Describe the methods used to ensure data quality in full detail, including how many peaks are at FDR 5% and above 5-fold enrichment.

Software

Describe the software used to collect and analyze the ChIP-seq data. For custom code that has been deposited into a community repository, provide accession details.

Flow Cytometry

Plots

Confirm that:

- ☐ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).
- ☐ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).
- ☐ All plots are contour plots with outliers or pseudocolor plots.
- ☐ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation

Describe the sample preparation, detailing the biological source of the cells and any tissue processing steps used.

Instrument

Identify the instrument used for data collection, specifying make and model number.

Software

Describe the software used to collect and analyze the flow cytometry data. For custom code that has been deposited into a community repository, provide accession details.

Cell population abundance

Describe the abundance of the relevant cell populations within post-sort fractions, providing details on the purity of the samples and how it was determined.

Gating strategy

Describe the gating strategy used for all relevant experiments, specifying the preliminary FSC/SSC gates of the starting cell population, indicating where boundaries between "positive" and "negative" staining cell populations are defined.

- ☐ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

Magnetic resonance imaging

Experimental design

Design type

Indicate task or resting state; event-related or block design.

Design specifications

Specify the number of blocks, trials or experimental units per session and/or subject, and specify the length of each trial or block (if trials are blocked) and interval between trials.

Behavioral performance measures

State number and/or type of variables recorded (e.g. correct button press, response time) and what statistics were used to establish that the subjects were performing the task as expected (e.g. mean, range, and/or standard deviation across subjects).

Acquisition

Imaging type(s)

Specify: functional, structural, diffusion, perfusion.

Field strength

Specify in Tesla

Sequence & imaging parameters

Specify the pulse sequence type (gradient echo, spin echo, etc.), imaging type (EPI, spiral, etc.), field of view, matrix size, slice thickness, orientation and TE/TR/flip angle.

Area of acquisition

State whether a whole brain scan was used OR define the area of acquisition, describing how the region was determined.

Diffusion MRI

☐ Used

☐ Not used

Preprocessing

Preprocessing software

Provide detail on software version and revision number and on specific parameters (model/functions, brain extraction, segmentation, smoothing kernel size, etc.).

Normalization

If data were normalized/standardized, describe the approach(es): specify linear or non-linear and define image types used for transformation OR indicate that data were not normalized and explain rationale for lack of normalization.

Normalization template

Describe the template used for normalization/transformation, specifying subject space or group standardized space (e.g. original Talairach, MNI305, ICBM152) OR indicate that the data were not normalized.

Noise and artifact removal

Describe your procedure(s) for artifact and structured noise removal, specifying motion parameters, tissue signals and physiological signals (heart rate, respiration).

Volume censoring

Define your software and/or method and criteria for volume censoring, and state the extent of such censoring.

Statistical modeling & inference

Model type and settings

Specify type (mass univariate, multivariate, RSA, predictive, etc.) and describe essential details of the model at the first and second levels (e.g. fixed, random or mixed effects; drift or auto-correlation).

Effect(s) tested

Define precise effect in terms of the task or stimulus conditions instead of psychological concepts and indicate whether ANOVA or factorial designs were used.

Specify type of analysis: ☐ Whole brain ☐ ROI-based ☐ BothStatistic type for inference
(See [Eklund et al. 2016](#))

Specify voxel-wise or cluster-wise and report all relevant parameters for cluster-wise methods.

Correction

Describe the type of correction and how it is obtained for multiple comparisons (e.g. FWE, FDR, permutation or Monte Carlo).

Models & analysis

n/a | Involved in the study

☐ ☐ Functional and/or effective connectivity☐ ☐ Graph analysis☐ ☐ Multivariate modeling or predictive analysis

Functional and/or effective connectivity

Report the measures of dependence used and the model details (e.g. Pearson correlation, partial correlation, mutual information).

Graph analysis

Report the dependent variable and connectivity measure, specifying weighted graph or binarized graph, subject- or group-level, and the global and/or node summaries used (e.g. clustering coefficient, efficiency, etc.).

Multivariate modeling and predictive analysis

Specify independent variables, features extraction and dimension reduction, model, training and evaluation metrics.

Genome-wide cell-free DNA fragmentation in patients with cancer

Stephen Cristiano^{1,2,15}, Alessandro Leal^{1,15}, Jillian Phallen^{1,15}, Jacob Fiksel^{1,2,15}, Vilmos Adleff¹, Daniel C. Bruhm¹, Sarah Østrup Jensen³, Jamie E. Medina¹, Carolyn Hruban¹, James R. White¹, Doreen N. Palsgrove¹, Noushin Niknafs¹, Valsamo Anagnostou¹, Patrick Forde¹, Jarushka Naidoo¹, Kristen Marrone¹, Julie Brahmer¹, Brian D. Woodward⁴, Hatim Husain⁴, Karlijn L. van Rooijen⁵, Mai-Britt Worm Ørntoft³, Anders Husted Madsen⁶, Cornelis J. H. van de Velde⁷, Marcel Verheij⁸, Annemieke Cats⁹, Cornelis J. A. Punt¹⁰, Geraldine R. Vink⁵, Nicole C. T. van Grieken¹¹, Miriam Koopman⁵, Remond J. A. Fijneman¹², Julia S. Johansen¹³, Hans Jørgen Nielsen¹⁴, Gerrit A. Meijer¹², Claus Lindbjerg Andersen³, Robert B. Scharpf^{1,2,*} & Victor E. Velculescu^{1,*}

Cell-free DNA in the blood provides a non-invasive diagnostic avenue for patients with cancer¹. However, characteristics of the origins and molecular features of cell-free DNA are poorly understood. Here we developed an approach to evaluate fragmentation patterns of cell-free DNA across the genome, and found that profiles of healthy individuals reflected nucleosomal patterns of white blood cells, whereas patients with cancer had altered fragmentation profiles. We used this method to analyse the fragmentation profiles of 236 patients with breast, colorectal, lung, ovarian, pancreatic, gastric or bile duct cancer and 245 healthy individuals. A machine learning model that incorporated genome-wide fragmentation features had sensitivities of detection ranging from 57% to more than 99% among the seven cancer types at 98% specificity, with an overall area under the curve value of 0.94. Fragmentation profiles could be used to identify the tissue of origin of the cancers to a limited number of sites in 75% of cases. Combining our approach with mutation-based cell-free DNA analyses detected 91% of patients with cancer. The results of these analyses highlight important properties of cell-free DNA and provide a proof-of-principle approach for the screening, early detection and monitoring of human cancer.

Much of the morbidity and mortality of human cancers worldwide results from late diagnosis when therapeutic intervention is less effective^{2,3}. Unfortunately, clinically proven biomarkers that can be used to broadly diagnose and treat patients are not widely available⁴. Recent analyses of circulating cell-free DNA (cfDNA) suggest that approaches using tumour-specific alterations may provide new opportunities for early diagnosis, but not all patients have detectable changes^{5–8}. Whole-genome sequencing (WGS) of cfDNA can identify chromosomal abnormalities in patients with cancer but detecting such alterations may be challenging owing to the small number of abnormal chromosomal changes^{9–12}. Analyses of the size of fragments of cfDNA have been contradictory, indicating both increases^{13–15} and decreases in the overall distribution of cfDNA^{12,16,17–19}. Recent studies have suggested that size selection of small cfDNA can increase enrichment of circulating tumour DNA in patients with late-stage cancer¹⁷. Nucleosome positions^{18,20}, patterns near transcription start sites^{20,21}, and the end positions of cfDNA²² may be altered in cancer, but the sequencing needed to identify nucleosomes is impractical for routine analyses.

Conceptually, the sensitivity of any cfDNA approach depends on the number of alterations examined as well as the technical and biological limitations of detecting such changes. As a typical blood sample contains approximately 2,000 genome equivalents of cfDNA per millilitre of plasma⁵, the theoretical limit of detection of a single alteration can be no better than one in a few thousand mutant to wild-type molecules. We hypothesized that the detection of a larger number of alterations in the genome may be more sensitive for detecting cancer in

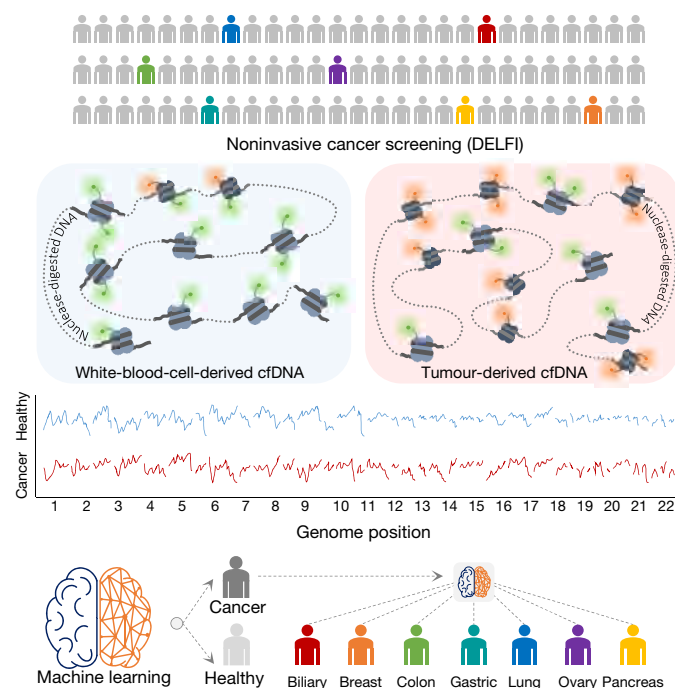


Fig. 1 | Schematic of DELFI approach. Blood is collected from healthy individuals and patients with cancer. cfDNA is extracted from plasma, processed into sequencing libraries, examined by WGS, mapped to the genome, and analysed to determine cfDNA fragmentation profiles across the genome. Machine learning is used to categorize whether individuals have cancer and identify the tumour tissue of origin.

¹The Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University School of Medicine, Baltimore, MD, USA. ²Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA. ³Department of Molecular Medicine, Aarhus University Hospital, Aarhus, Denmark. ⁴Division of Hematology and Oncology, Moores Cancer Center, University of California, San Diego, La Jolla, CA, USA. ⁵Department of Medical Oncology, University Medical Center, Utrecht University, Utrecht, The Netherlands. ⁶Department of Surgery, Herning Regional Hospital, Herning, Denmark. ⁷Department of Surgery, Leiden University Medical Center, Leiden, The Netherlands. ⁸Department of Radiation Oncology, The Netherlands Cancer Institute, Amsterdam, The Netherlands. ⁹Department of Gastrointestinal Oncology, The Netherlands Cancer Institute, Amsterdam, The Netherlands. ¹⁰Department of Medical Oncology, Academic Medical Center, University of Amsterdam, Amsterdam, The Netherlands. ¹¹Department of Pathology, VU University Medical Center, Amsterdam, The Netherlands. ¹²Department of Pathology, The Netherlands Cancer Institute, Amsterdam, The Netherlands. ¹³Department of Oncology, Herlev and Gentofte Hospital, Copenhagen University Hospital, Herlev, Denmark. ¹⁴Department of Surgical Gastroenterology 360, Hvidovre Hospital, Hvidovre, Denmark. ¹⁵These authors contributed equally: Stephen Cristiano, Alessandro Leal, Jillian Phallen, Jacob Fiksel. *e-mail: rscharpf@jhu.edu; velculescu@jhmi.edu

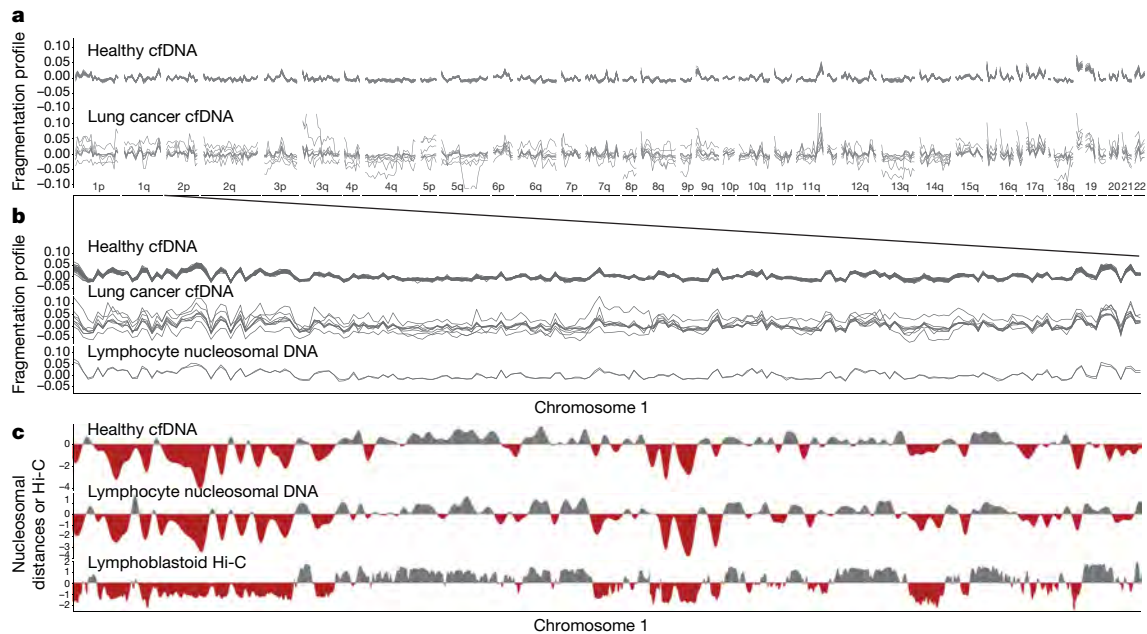


Fig. 2 | Aberrant cfDNA fragmentation profiles in patients with cancer. **a**, Genome-wide cfDNA fragmentation profiles (defined as the ratio of short to long fragments) from approximately $9\times$ WGS are shown in 5-Mb bins for 30 healthy individuals (top) and 8 patients with lung cancer (bottom). **b**, Analyses of healthy cfDNA (top), lung cancer cfDNA (middle), and healthy lymphocyte (bottom) fragmentation profiles from chromosome 1 at 1-Mb resolution. Healthy lymphocyte

profiles were scaled with a standard deviation equal to that of the median healthy cfDNA profiles. **c**, Smoothed median distances between adjacent nucleosomes centred at zero using 100 kb bins from healthy cfDNA (top) and nuclease-digested healthy lymphocytes (middle) are depicted together with the first eigenvector for the genome contact matrix from Hi-C analyses of lymphoblastoid cells²⁷ (bottom).

the circulation. Monte Carlo simulations showed that increasing the number of abnormalities detected from a few to tens or hundreds can improve the limit of detection, similar to recent analyses of methylation changes in cfDNA²³ (Extended Data Fig. 1a).

We developed an approach called 'DNA evaluation of fragments for early interception' (DELFI) (Fig. 1) to detect a large number of abnormalities in cfDNA by genome-wide analysis of fragmentation patterns. The method is based on low-coverage WGS of isolated cfDNA. Mapped sequences are analysed in non-overlapping windows that cover the genome. Conceptually, windows may range in size from thousands to millions of bases, resulting in hundreds to thousands of windows in the genome. We used 5-megabase (Mb) windows to evaluate cfDNA fragmentation patterns as this provided more than 20,000 reads per window at $1\text{--}2\times$ genome coverage. Within each window, we examined the coverage and size distribution of cfDNA fragments in healthy and cancer populations (Supplementary Table 1). The genome-wide pattern from an individual can be compared to reference populations to determine whether the pattern is likely to be healthy or cancer-derived. As genome-wide profiles may reveal differences associated with specific tissues, these patterns may also indicate the tissue source of cfDNA.

We focused on fragmentation size of cfDNA as we found that cancer-derived cfDNA may be more variable in length than cfDNA from non-cancer cells. We initially examined cfDNA from targeted regions captured and sequenced at high coverage from patients with breast, colorectal, lung or ovarian cancer⁵ (Supplementary Tables 1–3). Analyses of loci containing 165 tumour-specific alterations from 81 patients revealed an average absolute difference of 6.5 base pairs (bp; 95% confidence interval (CI), 5.4–7.6 bp) between the lengths of median mutant and wild-type cfDNA fragments, with mutant cfDNA fragments ranging from 30 bases smaller to 47 bases larger (Extended Data Fig. 1b, Supplementary Table 3). The GC content was similar for mutated and non-mutated fragments, with no correlation between GC content and fragment length (Extended Data Fig. 1c, d). Analyses of 44 germline alterations from 38 patients identified median cfDNA size differences of less than 1 bp between different alleles (Extended Data Fig. 2a, Supplementary Table 3). For 41 alterations related to

clonal haematopoiesis⁵, there were no significant differences between cfDNA fragments containing such alterations and wild-type fragments (Extended Data Fig. 2b, Supplementary Table 3). Overall, the lengths of cancer-derived cfDNA fragments were more variable than non-cancer cfDNA ($P < 0.001$, variance ratio test). We hypothesized that these differences may reflect changes in chromatin structure as well as other genomic and epigenomic abnormalities in cancer^{24,25}, and that cfDNA fragmentation in a position-specific manner could serve as a biomarker for cancer detection.

As targeted sequencing analyses a limited number of loci, we investigated whether genome-wide analyses would detect additional abnormalities from cfDNA fragmentation. In a pilot analysis, we isolated cfDNA from around 4 ml of plasma from 8 patients with stage I–III lung cancer and 30 healthy individuals (Supplementary Tables 1, 4, 5), and performed WGS at approximately $9\times$ coverage (Supplementary Table 4). As expected^{12,18,19}, the median overall lengths of fragments of cfDNA from healthy individuals were larger than those from patients with cancer (167.3 bp and 163.8, respectively, $P < 0.01$, Welch's *t*-test) (Supplementary Table 5). To examine differences in fragment size and coverage in a position-dependent manner across the genome, we mapped fragments to their genomic origin and evaluated fragment lengths in 504 windows of 5 Mb, covering approximately 2.6 Gb of the genome. For each window, we determined the fraction of small cfDNA fragments (100–150 bp) to larger cfDNA fragments (151–220 bp) and overall coverage to obtain genome-wide fragmentation profiles for each sample.

We found that healthy individuals had similar genome-wide fragmentation profiles (Fig. 2a, b, Extended Data Fig. 3a). To examine the origins of cfDNA fragmentation patterns, we isolated and nuclease-treated nuclei from lymphocytes of two healthy individuals to obtain nucleosomal DNA fragments. Healthy cfDNA patterns were highly correlated to lymphocyte nucleosomal DNA fragmentation profiles and nucleosome distances (Fig. 2b, c, Extended Data Fig. 3b, c). Median distances between nucleosomes in lymphocytes were correlated to high-throughput sequencing chromosome conformation capture (Hi-C) open (A) and closed (B) compartments of lymphoblastoid cells^{26,27}

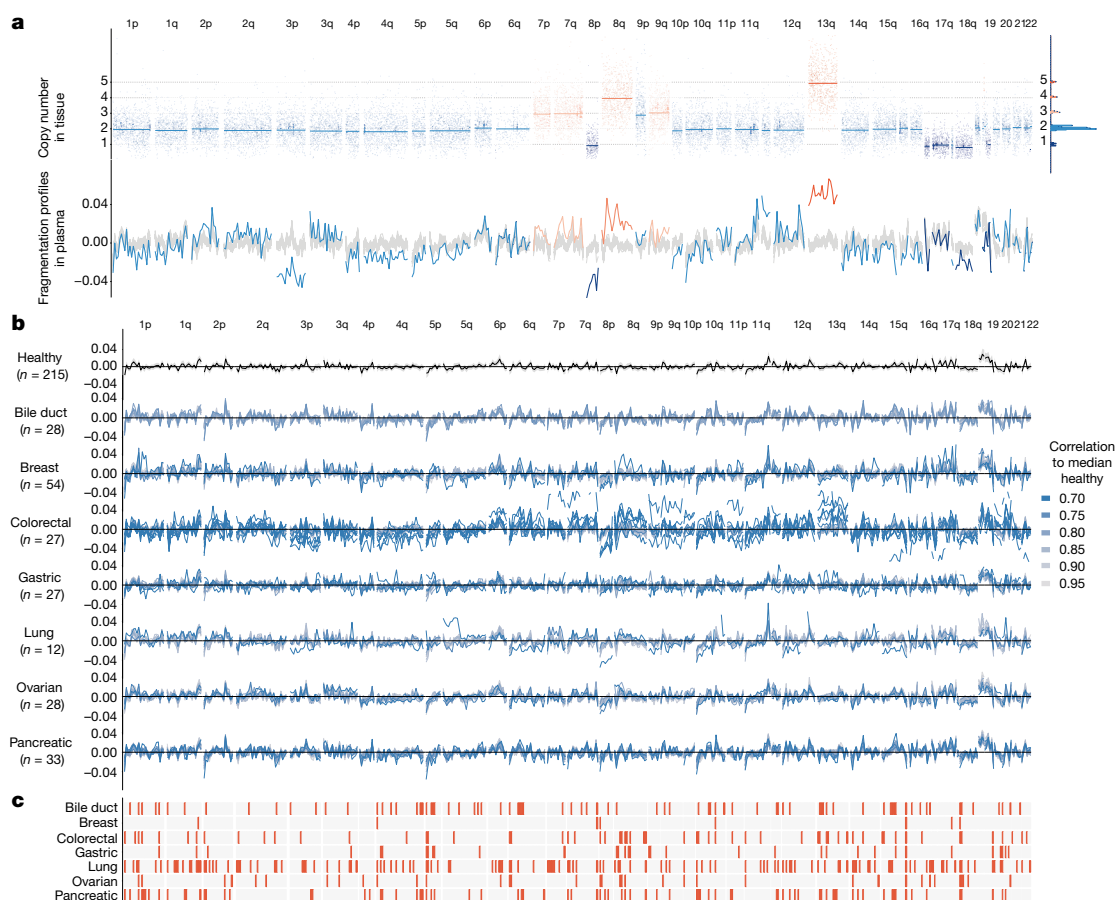


Fig. 3 | cfDNA fragmentation profiles in healthy individuals and patients with cancer. **a**, Fragmentation profiles (bottom) in the context of tumour copy number changes (top) in a patient with colorectal cancer. The distribution of segment means and integer copy numbers are shown at top right. **b**, GC-adjusted fragmentation profiles from WGS at 1–2× coverage for healthy individuals and patients with cancer are depicted per

cancer type using 5-Mb windows. The median healthy profile is indicated in black and the 98% confidence band is shown in grey. For patients with cancer, individual profiles are coloured based on their Pearson correlation to the healthy median. **c**, Windows are indicated in orange if more than 10% of the cancer samples had a fragment ratio more than three standard deviations from the median healthy fragment ratio.

(Fig. 2c). These analyses suggest that fragmentation patterns of normal cfDNA are the result of nucleosomal DNA patterns that reflect the chromatin structure of normal blood cells.

In contrast to healthy cfDNA, patients with cancer had several distinct genomic differences with increases and decreases in fragment sizes at different regions (Fig. 2a, b). We performed genome-wide correlation analyses of the fraction of short to long cfDNA fragments for each sample compared to the median fragment length profile of healthy individuals, and found that—although cfDNA fragment profiles were consistent among healthy individuals (median correlation of 0.99)—the median correlation of fragment ratios among patients with cancer was 0.84 ($P < 0.001$, Wilcoxon rank-sum test; Fig. 2a, b, Extended Data Fig. 3d, Supplementary Table 5). Similar differences were observed when comparing cfDNA fragmentation profiles of patients with cancer to fragmentation profiles of healthy lymphocytes (Fig. 2c, Extended Data Fig. 3b, c). To account for potential biases attributable to GC content, we applied a locally weighted smoother and found that differences in fragmentation profiles between healthy individuals and patients with cancer remained after this adjustment (median correlation of patients with cancer to healthy = 0.83, Supplementary Table 5).

We subsampled WGS data at 9× coverage to approximately 2×, 1×, 0.5×, 0.2× and 0.1× genome coverage, and determined that altered fragmentation profiles from patients with cancer were identified even at 0.5× coverage (Extended Data Fig. 3e, f). On the basis of these observations, we performed WGS at 1–2× coverage to evaluate whether fragmentation profiles may change during the course of therapy^{28,29}. We evaluated cfDNA from 19 patients with non-small-cell lung cancer

during therapy with anti-EGFR or anti-ERBB2 agents (Supplementary Table 6). The degree of abnormality in the fragmentation profiles during therapy closely matched levels of EGFR or ERBB2 mutant allele fractions²⁹ (Extended Data Fig. 4, Spearman correlation of mutant allele fractions to fragmentation profiles = 0.74). These results demonstrate that fragmentation analyses may be useful for detecting tumour-derived cfDNA and monitoring patients during treatment.

As cfDNA fragmentation profiles would be expected to reflect both epigenomic and genomic alterations, we examined these in a patient with known tumour copy number changes. Altered fragmentation profiles were present in regions of the genome that were copy-neutral and were further affected in regions with copy number changes (Fig. 3a, Extended Data Fig. 5a). Position-dependent differences in fragmentation patterns distinguished cancer-derived cfDNA from healthy cfDNA, whereas analyses of overall fragment sizes of cfDNA would have missed such differences (Extended Data Fig. 5a, b).

We performed WGS at 1–2× coverage of cfDNA from 208 patients with cancer, including breast ($n = 54$), colorectal ($n = 27$), lung ($n = 12$), ovarian ($n = 28$), pancreatic ($n = 34$), gastric ($n = 27$) or bile duct cancer ($n = 26$), as well as 215 healthy individuals (Supplementary Tables 1, 4). All patients with cancer had not undergone previous treatment and most had resectable disease ($n = 183$). After GC adjustment of short and long cfDNA fragment coverage (Extended Data Fig. 6a, b), we examined coverage and size characteristics of fragments in windows throughout the genome (Fig. 3b, Supplementary Tables 4, 7). Healthy individuals had concordant fragmentation profiles whereas patients with cancer had highly variable profiles with decreased correlation to the median healthy profile (Supplementary Table 7). An

Table 1 | DELFI performance for cancer detection

Individuals analysed			95% specificity			98% specificity		
			Individuals detected	Sensitivity (%)	95% CI (%)	Individuals detected	Sensitivity (%)	95% CI (%)
Healthy		215	10	—	—	4	—	—
Cancer		208	166	80	74–85	152	73	67–79
Type	Breast	54	38	70	56–82	31	57	43–71
	Bile duct	26	23	88	70–98	21	81	61–93
	Colorectal	27	22	81	62–94	19	70	50–86
	Gastric	27	22	81	62–94	22	81	62–94
	Lung	12	12	100	74–100	12	100	74–100
	Ovarian	28	25	89	72–98	25	89	72–98
	Pancreatic	34	24	71	53–85	22	65	46–80
Stage	I	41	30	73	53–86	28	68	52–82
	II	109	85	78	69–85	78	72	62–80
	III	33	30	91	76–98	26	79	61–91
	IV	22	18	82	60–95	17	77	55–92
	X	3	3	100	29–100	3	100	29–100

analysis of commonly altered genomic windows revealed a median of 60 affected windows across the cancer types analysed, which highlights position-dependent alterations in fragmentation of cfDNA (Fig. 3c).

We implemented a gradient tree boosting machine learning model to examine whether cfDNA has characteristics of a patient with cancer or healthy individual, and estimated performance characteristics of this approach by tenfold cross-validation repeated ten times (Extended Data Fig. 7a, b). The machine learning model included GC-adjusted short and long fragment coverage characteristics in windows throughout the genome. We also developed a machine learning classifier for copy number changes from chromosomal arm features^{10,11} (Extended Data Fig. 8a, Supplementary Table 8) and included mitochondrial copy number changes¹² (Extended Data Fig. 8b). Using this implementation of DELFI, we obtained a score that could be used to classify patients as being healthy or having cancer. We detected 152 out of 208 cancer patients (73% sensitivity, 95% confidence interval 67–79%), and misclassified 4 out of 215 healthy individuals (98% specificity) (Table 1). At a threshold of 95% specificity, we detected 80% of patients with cancer (95% confidence interval 74–85%), including 79% of patients with resectable (stage I–III) disease (145 out of 183) and 82% of patients with stage IV disease (18 of 22) (Table 1). Receiver operator characteristic analyses for the detection of patients with cancer had an area under the curve (AUC) value of 0.94 (95% confidence interval 0.92–0.96), ranging from 0.86 for pancreatic cancer to at least 0.93 for breast, bile duct, colorectal, gastric, lung and ovarian cancers (Fig. 4, Extended Data Fig. 9a), with AUC values of at least 0.92 for each stage (Extended Data Fig. 9b). To assess the contribution of fragment size and coverage across the genome, chromosome arm copy number or mitochondrial copy number to the predictive accuracy of the model, we implemented the cross-validation procedure to assess performance characteristics of these features in isolation. Fragment coverage features alone (AUC = 0.94) were nearly identical to the classifier that combined all features (AUC = 0.94). By contrast, machine learning analyses of changes in chromosomal copy number had lower performance (AUC = 0.88) but were still more predictive than copy number using individual scores (AUC = 0.78) or mitochondrial copy number (AUC = 0.72) (Fig. 4). These results suggest that fragment coverage is the major contributor to our classifier, but we have included all features in our prediction model as they can be obtained from the same WGS data and may contribute in a complementary fashion for cancer detection.

As fragmentation profiles reveal regional differences between tissues, we used machine learning to identify the tissue of origin of circulating tumour DNA. These analyses had a 61% accuracy (95% confidence

interval 53–67%) that increased to 75% (95% confidence interval 69–81%) when assigning circulating tumour DNA to one of two sites of origin (Extended Data Fig. 9c, d). For all tumour types, the classification of tissue of origin by DELFI was higher than that by random assignment ($P < 0.01$, binomial test, Extended Data Fig. 9d).

We evaluated whether combining DELFI with mutation detection in cfDNA⁵ could increase the sensitivity of cancer detection (Extended Data Fig. 10). An evaluation of cases analysed using both approaches revealed that 82% (103 out of 126) of patients were detected using DELFI, and 66% (83 out of 126) had sequence alterations. For cases with mutant allele fractions of less than 1%, DELFI detected 80% of cases—including those that were undetectable using targeted sequencing (Supplementary Table 7). When these approaches were used together, the combined sensitivity increased to 91% (115 out of 126 patients) with a specificity of 98% (Extended Data Fig. 10).

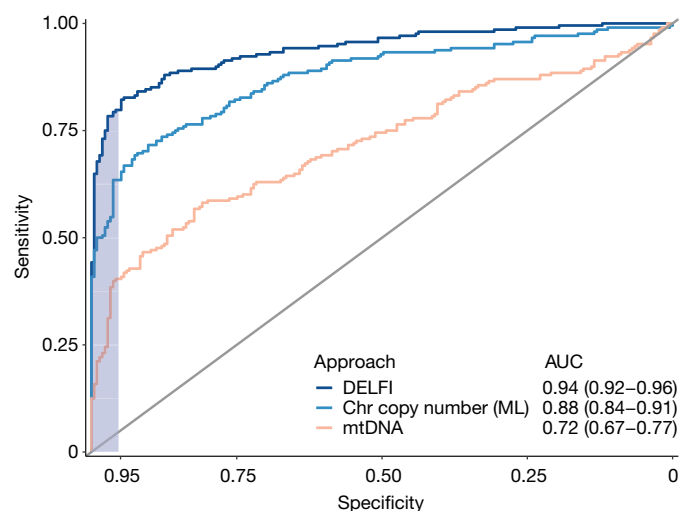


Fig. 4 | Detection of cancer using DELFI. Receiver operator characteristics for the detection of cancer using cfDNA fragmentation profiles and other genome-wide features in a machine learning approach are depicted for a cohort of 215 healthy individuals and 208 patients with cancer (DELFI, AUC = 0.94), with $\geq 95\%$ specificity shaded in blue. Machine learning analyses of chromosomal arm copy number (Chr copy number (ML)), and mitochondrial genome copy number analyses (mtDNA) are shown.

Overall, we have determined that genome-wide fragmentation profiles of cfDNA are different between patients with cancer and healthy individuals. In patients with cancer, fragmentation patterns in cfDNA appear to result from mixtures of nucleosomal DNA from both blood and neoplastic cells. Our approach could be further improved through recovery of smaller fragments^{17,30}, evaluation of single-stranded libraries^{18,30,31} or use of alternative technologies. Additionally, PCR-free libraries could reduce GC bias and sequencing artefacts^{18,30,31}.

These observations have important implications for non-invasive detection of human cancer. DELFI simultaneously analyses tens to hundreds of tumour-specific abnormalities from minute amounts of cfDNA, overcoming a limitation that has precluded the possibility of more-sensitive analyses of cfDNA. These analyses detected a higher fraction of patients with cancer than previous methods^{5–7,12,17}, and combining DELFI with the detection of sequence alterations in cfDNA further increased the sensitivity of detection. As fragmentation profiles seem to be related to nucleosomal patterns, DELFI may be useful for determining the source of tumour-derived cfDNA, an aspect that could be further improved using clinical characteristics, methylation changes²³ and other diagnostic approaches⁶. DELFI requires only a small amount of whole-genome sequencing, which suggests that this approach could be broadly applied for the screening and management of patients with cancer.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at <https://doi.org/10.1038/s41586-019-1272-6>.

Received: 19 November 2018; Accepted: 10 May 2019;

Published online 29 May 2019.

- Wan, J. C. M. et al. Liquid biopsies come of age: towards implementation of circulating tumour DNA. *Nat. Rev. Cancer* **17**, 223–238 (2017).
- Bray, F. et al. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **68**, 394–424 (2018).
- World Health Organization. *Guide to Cancer Early Diagnosis* https://www.who.int/cancer/publications/cancer_early_diagnosis/en/ (WHO, 2017).
- National Comprehensive Cancer Network. *NCCN Clinical Practice Guidelines in Oncology* https://www.nccn.org/professionals/physician_gls/default.aspx (accessed 16 April 2019).
- Phallen, J. et al. Direct detection of early-stage cancers using circulating tumor DNA. *Sci. Transl. Med.* **9**, eaan2415 (2017).
- Cohen, J. D. et al. Detection and localization of surgically resectable cancers with a multi-analyte blood test. *Science* **359**, 926–930 (2018).
- Newman, A. M. et al. An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage. *Nat. Med.* **20**, 548–554 (2014).
- Bettegowda, C. et al. Detection of circulating tumor DNA in early- and late-stage human malignancies. *Sci. Transl. Med.* **6**, 224ra24 (2014).
- Leary, R. J. et al. Development of personalized tumor biomarkers using massively parallel sequencing. *Sci. Transl. Med.* **2**, 20ra14 (2010).
- Leary, R. J. et al. Detection of chromosomal alterations in the circulation of cancer patients with whole-genome sequencing. *Sci. Transl. Med.* **4**, 162ra154 (2012).
- Chan, K. C. et al. Noninvasive detection of cancer-associated genome-wide hypomethylation and copy number aberrations by plasma DNA bisulfite sequencing. *Proc. Natl Acad. Sci. USA* **110**, 18761–18768 (2013).
- Jiang, P. et al. Lengthening and shortening of plasma DNA in hepatocellular carcinoma patients. *Proc. Natl Acad. Sci. USA* **112**, E1317–E1325 (2015).
- Wang, B. G. et al. Increased plasma DNA integrity in cancer patients. *Cancer Res.* **63**, 3966–3968 (2003).
- Umetani, N. et al. Prediction of breast tumor progression by integrity of free circulating DNA in serum. *J. Clin. Oncol.* **24**, 4270–4276 (2006).
- Chan, K. C., Leung, S. F., Yeung, S. W., Chan, A. T. & Lo, Y. M. Persistent aberrations in circulating DNA integrity after radiotherapy are associated with poor prognosis in nasopharyngeal carcinoma patients. *Clin. Cancer Res.* **14**, 4141–4145 (2008).
- Mouliere, F. et al. High fragmentation characterizes tumour-derived circulating DNA. *PLoS ONE* **6**, e23418 (2011).
- Mouliere, F. et al. Enhanced detection of circulating tumor DNA by fragment size analysis. *Sci. Transl. Med.* **10**, eaat4921 (2018).
- Snyder, M. W., Kircher, M., Hill, A. J., Daza, R. M. & Shendure, J. Cell-free DNA comprises an *in vivo* nucleosome footprint that informs its tissues-of-origin. *Cell* **164**, 57–68 (2016).
- Underhill, H. R. et al. Fragment length of circulating tumor DNA. *PLoS Genet.* **12**, e1006162 (2016).
- Ulz, P. et al. Inferring expressed genes by whole-genome sequencing of plasma DNA. *Nat. Genet.* **48**, 1273–1278 (2016).
- Ivanov, M., Baranova, A., Butler, T., Spellman, P. & Mileyko, V. Non-random fragmentation patterns in circulating cell-free DNA reflect epigenetic regulation. *BMC Genomics* **16** (Suppl. 13), S1 (2015).
- Jiang, P. et al. Preferred end coordinates and somatic variants as signatures of circulating tumor DNA associated with hepatocellular carcinoma. *Proc. Natl Acad. Sci. USA* **115**, E10925–E10933 (2018).
- Shen, S. Y. et al. Sensitive tumour detection and classification using plasma cell-free DNA methylomes. *Nature* **563**, 579–583 (2018).
- Corces, M. R. et al. The chromatin accessibility landscape of primary human cancers. *Science* **362**, eaav1898 (2018).
- Polak, P. et al. Cell-of-origin chromatin organization shapes the mutational landscape of cancer. *Nature* **518**, 360–364 (2015).
- Lieberman-Aiden, E. et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).
- Fortin, J. P. & Hansen, K. D. Reconstructing A/B compartments as revealed by Hi-C using long-range correlations in epigenetic data. *Genome Biol.* **16**, 180 (2015).
- Diehl, F. et al. Circulating mutant DNA to assess tumor dynamics. *Nat. Med.* **14**, 985–990 (2008).
- Phallen, J. et al. Early noninvasive detection of response to targeted therapy in non-small cell lung cancer. *Cancer Res.* **79**, 1204–1213 (2019).
- Burnham, P. et al. Single-stranded DNA library preparation uncovers the origin and diversity of ultrashort cell-free DNA in plasma. *Sci. Rep.* **6**, 27859 (2016).
- Sanchez, C., Snyder, M. W., Tanos, R., Shendure, J. & Thierry, A. R. New insights into structural features and optimal detection of circulating tumor DNA determined by single-strand DNA analysis. *NPL Genom. Med.* **3**, 31 (2018).

Acknowledgements We thank members of our laboratories for critical review of the manuscript. This work was supported, in part, by the Dr. Miriam and Sheldon G. Adelson Medical Research Foundation, the Stand Up to Cancer–Dutch Cancer Society International Translational Cancer Research Dream Team Grant (SU2C-AACR-DT1415), the Commonwealth Foundation, the Cigarette Restitution Fund, the Burroughs Wellcome Fund and the Maryland Genetics, Epidemiology and Medicine Training Program, the AACR-Janssen Cancer Interception Research Fellowship, the Mark Foundation for Cancer Research, US NIH (grants CA121113, CA006973, and CA180950), the Danish Council for Independent Research (11-105240), the Danish Council for Strategic Research (1309-00006B), the Novo Nordisk Foundation (NNF14OC0012747 and NNF17OC0025052), and the Danish Cancer Society (R133-A8520-00-S41 and R146-A9466-16-S2). Stand Up To Cancer is a program of the Entertainment Industry Foundation administered by the American Association for Cancer Research.

Reviewer information Nature thanks Daniel De Carvalho, Ellen Heitzer and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions S.C., A.L., J.P., J.F., V. Adleff, R.B.S. and V.E.V. designed and planned the study, and developed and optimized experimental protocols. A.L., J.P., V. Adleff, J.E.M. and D.N.P. performed experiments. S.Ø.J., V. Anagnostou, P.F., J.N., K.M., J.B., B.D.W., H.H., K.L.V.R., M.-B.W.Ø., A.H.M., C.J.H.v.d.V., M.V., A.C., C.J.A.P., G.R.V., N.C.T.v.G., M.K., R.J.A.F., J.S.J., H.J.N., G.A.M. and C.L.A. organized patient enrolment, sample collection, and clinical data curation. S.C., A.L., J.P., J.F., V. Adleff, D.C.B., J.E.M., J.R.W., N.N., G.A.M., C.L.A., R.B.S. and V.E.V. analysed and interpreted data. S.C., A.L., J.P., J.F., R.B.S. and V.E.V. wrote the manuscript and incorporated feedback from all authors. S.C., A.L., J.P. and J.F. contributed equally to this study.

Competing interests S.C., A.L., J.P., J.F., V. Adleff, R.B.S. and V.E.V. are inventors on patent applications (62/673,516 and 62/795,900) submitted by Johns Hopkins University related to cell-free DNA for cancer detection. V.E.V. is a founder of Delfi Diagnostics and Personal Genome Diagnostics, a member of their Scientific Advisory Boards and Boards of Directors, and owns Delfi Diagnostics and Personal Genome Diagnostics stock, which are subject to certain restrictions under university policy. Within the last five years, V.E.V. has been an advisor to Daiichi Sankyo, Janssen Diagnostics, Ignyta, and Takeda Pharmaceuticals. The terms of these arrangements are managed by Johns Hopkins University in accordance with its conflict of interest policies.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-019-1272-6>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-019-1272-6>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to R.B.S. or V.E.V.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

METHODS

Patient and sample characteristics. Plasma samples from healthy individuals and plasma and tissue samples from patients with breast, lung, ovarian, colorectal, bile duct or gastric cancer were obtained from ILSBio/Bioreclamation, Aarhus University, Herlev Hospital of the University of Copenhagen, Hvidovre Hospital, the University Medical Center of the University of Utrecht, the Academic Medical Center of the University of Amsterdam, the Netherlands Cancer Institute and the University of California, San Diego. All samples were obtained under Institutional Review Board approved protocols with informed consent from all participants for research use at participating institutions. Plasma samples from healthy individuals were obtained at the time of routine screening, including for colonoscopies or Pap smears. Individuals were considered healthy if they had no previous history of cancer and negative screening results.

Plasma samples from individuals with breast, colorectal, gastric, lung, ovarian, pancreatic and bile duct cancer were obtained at the time of diagnosis, before tumour resection or therapy. Nineteen patients with lung cancer analysed for changes in cfDNA fragmentation profiles across several time points were undergoing treatment with anti-EGFR or anti-ERBB2 therapy²⁹. Clinical data for all patients included in this study are listed in Supplementary Table 1. Sex was confirmed by genomic analyses of X and Y chromosome representation. Pathological staging of patients with gastric cancer was performed after neoadjuvant therapy. Samples for which the tumour stage was unknown were indicated as stage X.

Nucleosomal DNA purification. Viable frozen lymphocytes were elutriated from leukocytes obtained from a healthy male (C0618) and female (D0808-L) (Advanced Biotechnologies). Aliquots of 1×10^6 cells were used for nucleosomal DNA purification using EZ Nucleosomal DNA Prep Kit (Zymo Research). Cells were initially treated with 100 μ l of Nuclei Prep Buffer and incubated on ice for 5 min. After centrifugation at 200g for 5 min, supernatant was discarded and pelleted nuclei were treated twice with 100 μ l of Atlantis Digestion Buffer. Finally, cellular nucleic DNA was fragmented with 0.5 U of Atlantis dsDNase at 42 °C for 20 min. Reactions were stopped using 5 \times Stop Buffer and DNA was purified using Zymo-Spin IIC Columns. Concentration and quality of eluted cellular nucleic DNA were analysed using the Bioanalyzer 2100 (Agilent Technologies).

Sample preparation and sequencing of cfDNA. Whole blood was collected in EDTA tubes and processed immediately or within one day after storage at 4 °C, or was collected in Streck tubes and processed within two days of collection for three patients with cancer who were part of the monitoring analysis. Plasma and cellular components were separated by centrifugation at 800g for 10 min at 4 °C. Plasma was centrifuged a second time at 18,000g at room temperature to remove any remaining cellular debris and stored at –80 °C until the time of DNA extraction. DNA was isolated from plasma using the Qiagen Circulating Nucleic Acids Kit (Qiagen GmbH) and eluted in LoBind tubes (Eppendorf AG). Concentration and quality of cfDNA were assessed using the Bioanalyzer 2100 (Agilent Technologies).

Next-generation sequencing (NGS) cfDNA libraries were prepared for WGS and targeted sequencing using 5–250 ng of cfDNA as previously described⁵. In brief, genomic libraries were prepared using the NEBNext DNA Library Prep Kit for Illumina (New England Biolabs (NEB)) with four main modifications to the manufacturer's guidelines: (i) the library purification steps used the on-bead AMPure XP approach to minimize sample loss during elution and tube transfer steps³²; (ii) NEBNext End Repair, A-tailing and adaptor ligation enzyme and buffer volumes were adjusted as appropriate to accommodate the on-bead AMPure XP purification strategy; (iii) a pool of eight unique Illumina dual index adaptors with 8-bp barcodes was used in the ligation reaction; and (iv) cfDNA libraries were amplified with Phusion Hot Start Polymerase. Whole-genome libraries were sequenced using 100-bp paired-end runs on the Illumina HiSeq 2000/2500 (Illumina).

Analyses of targeted sequencing data from cfDNA. Analyses of targeted NGS data for cfDNA samples were performed as previously described⁵. In brief, primary processing was completed using Illumina Consensus Assessment of Sequence and Variation (CASAVA) software (v.1.8), including demultiplexing and masking of dual-index adaptor sequences. Sequence reads were aligned against the human reference genome (version hg18 or hg19) using NovoAlign with additional realignment of select regions using the Needleman–Wunsch method³³. The positions of sequence alterations we identified have not been affected by the different genome builds. Candidate mutations, consisting of point mutations, small insertions and deletions, were identified using VariantDx³³ (Personal Genome Diagnostics) across the targeted regions of interest.

To analyse the fragment lengths of cfDNA molecules, we required that each read pair from a cfDNA molecule had a Phred quality score ≥ 30 . We removed all duplicate DNA fragments, defined as having the same start, end and index barcode. For each mutation, we included only fragments for which one or both of the read pairs contained the mutated (or wild-type) base at the given position. This analysis was done using the R packages Rsamtools and GenomicAlignments.

For each genomic locus in which a somatic mutation was identified, we compared the lengths of fragments containing the mutant allele to the lengths of

fragments with the wild-type allele. If more than 100 mutant fragments were identified, we used Welch's two-sample *t*-test to compare the mean fragment lengths. For loci with fewer than 100 mutant fragments, we implemented a bootstrap procedure. Specifically, we sampled with replacement *N* fragments containing the wild-type allele, in which *N* denotes the number of fragments with the mutation. For each bootstrap replicate of wild-type fragments, we computed their median length. The *P* value was estimated as the fraction of bootstrap replicates with a median wild-type fragment length as long as, or more extreme than, the observed median mutant fragment length.

Analyses of WGS data from cfDNA. Primary processing of whole-genome NGS data for cfDNA samples was performed using Illumina CASAVA (Consensus Assessment of Sequence and Variation) software (v.1.8.2), including demultiplexing and masking of dual-index adaptor sequences. Sequence reads were aligned against the human reference genome (version hg19) using ELAND.

Read pairs with a MAPQ score below 30 for either read and PCR duplicates were removed. We tiled the hg19 autosomes into 26,236 adjacent, non-overlapping 100-kb bins. We excluded regions of low mappability based on previous work²⁷ in which 10% of bins with the lowest coverage were removed, and excluded reads that fell in the Duke blacklisted regions (<http://hgdownload.cse.ucsc.edu/goldenpath/hg19/encodeDCC/wgEncodeMapability/>). Using this approach, we excluded 361 Mb (13%) of the hg19 reference genome, including centromeric and telomeric regions. Short fragments were defined as having lengths between 100 and 150 bp and long fragments as having lengths between 151 and 220 bp.

To account for biases in coverage attributable to GC content of the genome, we applied locally weighted scatterplot smoothing (LOWESS, also known as LOESS) regression analysis with a span setting of 0.75 to the scatterplot of average fragment GC versus coverage calculated for each 100-kb bin. This LOESS regression was performed separately for short and long fragments to account for possible differences in GC effects on coverage in plasma by fragment length—an approach loosely motivated by a previous study³⁴. We subtracted the predictions for short and long coverage explained by GC from the LOESS model, obtaining residuals for short and long that were uncorrelated with GC. We returned the residuals to the original scale by adding back the genome-wide median short and long estimates of coverage. This procedure was repeated for each sample to account for possible differences in GC effects on coverage between samples. To reduce the feature space and noise further, we calculated the total GC-adjusted coverage in 5-Mb bins.

To compare the variability of fragment lengths from healthy subjects to fragments in patients with cancer, we calculated the standard deviation of the short to long fragmentation profiles for each individual. We compared the median of the standard deviations in the two groups by a Wilcoxon rank-sum test.

Analyses of changes in chromosome-arm copy number. To develop arm-level statistics for copy number changes, we adapted a previously described approach for aneuploidy detection in plasma, which used both chromosome-arm-specific Z-scores as well as plasma aneuploidy (PA) scores to summarize arm-level data¹⁰. This adapted approach divides the genome into non-overlapping 50-kb bins for which GC-corrected log₂-transformed read depth was obtained after correction by LOESS with span setting of 0.75. This LOESS-based correction is comparable to the approach outlined above, but is evaluated on a log₂ scale to increase robustness to outliers in the smaller bins and does not stratify by fragment length. To obtain an arm-specific Z-score for changes in copy number, the mean GC-adjusted read depth for each arm was centred and scaled by the mean and standard deviation, respectively, of read depth scores obtained from an independent set of 50 healthy samples.

Analyses of mitochondrial-aligned reads from cfDNA. Whole-genome sequence reads that initially mapped to the mitochondrial genome were extracted from .bam files and realigned to the hg19 reference genome in end-to-end mode with Bowtie2 as previously described³⁵. The resulting aligned reads were filtered such that both mates aligned to the mitochondrial genome with MAPQ ≥ 30 . The number of fragments mapping to the mitochondrial genome was counted and converted to a percentage of the total number of fragments in the original .bam files.

Prediction model for cancer detection. To distinguish healthy individuals from patients with cancer using fragmentation profiles, we used a stochastic gradient boosting model (gbm)^{36,37}. GC-corrected total and short fragment coverage for all 504 bins were centred and scaled for each sample to have mean zero and unit standard deviation. Additional features included Z-scores for each of the 39 autosomal arms and mitochondrial representation (log₁₀-transformed proportion of reads mapped to the mitochondria). To estimate the prediction error of this approach, we used tenfold cross-validation³⁸. Feature selection, performed only on the training data in each cross-validation run, removed bins that were highly correlated (correlation > 0.9) or had near-zero variance. Stochastic gradient boosted machine learning was implemented using the R package gbm with parameters: n.trees = 150, interaction.depth = 3, shrinkage = 0.1, and n.minobsinnode = 10. To average over the prediction error from the randomization of individuals to folds, we repeated the tenfold cross-validation procedure ten times. Confidence intervals

for sensitivity were obtained from 2,000 bootstrap replicates with specificity fixed at 98% and 95%.

Prediction model for tumour tissue of origin classification. For samples correctly identified from patients with cancer at 90% specificity ($n = 174$), a separate stochastic gradient boosting model was trained to classify the tissue of origin. To account for the small number of lung samples used for prediction, we included 18 cfDNA baseline samples from patients with late-stage lung cancer from the monitoring analyses of our study. Performance characteristics of the model were evaluated using tenfold cross-validation repeated ten times. This gbm model was trained using the same features as in the cancer classification model. Features that displayed correlation above 0.9 to each other or had near zero variance were removed within each training dataset during cross-validation. The tissue class probabilities were averaged across the ten replicates for each patient and the class with the highest probability was used as the predicted tissue.

Analyses of nucleosomal DNA from human lymphocytes and cfDNA. From the nuclease-treated lymphocytes, fragment sizes were analysed in 5-Mb bins as described for whole-genome cfDNA analyses. A genome-wide map of nucleosome positions was constructed from the nuclease-treated lymphocyte cell lines. This approach identified local biases in the coverage of circulating fragments, indicating a region protected from degradation. A 'window positioning score' (WPS) was used to score each base pair in the genome¹⁸. Using a sliding window of 60-bp centred around each base, the WPS was calculated as the number of fragments completely spanning the window minus the number of fragments with only one end in the window. Because fragments arising from nucleosomes have a median length of 167 bp, a high WPS indicated a possible nucleosomal position. WPS values were centred at zero using a running median and smoothed using a Kolmogorov–Zurbenko filter³⁹. For spans of positive WPS between 50 and 450 bp, a nucleosome peak was defined as the set of base pairs with a WPS above the median in that window. The calculation of nucleosome positions for cfDNA from 30 healthy individuals with sequence coverage of $9\times$ was determined in the same manner as for lymphocyte DNA. To ensure that nucleosomes in healthy cfDNA were representative, we defined a consensus track of nucleosomes consisting only of nucleosomes identified in two or more individuals. Median distances between adjacent nucleosomes were calculated from the consensus track.

Monte Carlo simulation of detection sensitivity. We used Monte Carlo simulation to estimate the probability of detecting a molecule with a tumour-derived alteration. In brief, we generated one million molecules from a multinomial distribution. For a simulation with m alterations, wild-type molecules were simulated with probability p and each of the m tumour alterations were simulated with probability $(1 - p)/m$. Next, we sampled $g \times m$ molecules randomly with replacement, in which g denotes the number of genome equivalents in 1 ml of plasma. If a tumour alteration was sampled (s) or more times, we classified the sample as cancer-derived. We repeated the simulation 1,000 times, estimating the probability that the

in silico sample would be correctly classified as cancer by the mean of the cancer indicator. Setting $g = 2,000$ and $s = 5$, we varied the number of tumour alterations by powers of 2 from 1 to 256 and the fraction of tumour-derived molecules from 0.0001% to 1%.

Statistical analyses. All statistical analyses were performed using R version 3.4.3. The R packages caret (v.6.0-79) and gbm (v.2.1-4) were used to implement the classification of healthy versus cancer and tissue of origin. Confidence intervals from the model output were obtained with the pROC (v.1.13) R package⁴⁰. Assuming the prevalence of undiagnosed cancer cases in this population is high (1 or 2 cases per 100 healthy), a genomic assay with a specificity of 0.95 and sensitivity of 0.8 would have useful operating characteristics (positive predictive value of 0.25 and negative predictive value near 1). Power calculations suggest that an analysis of more than 200 patients with cancer and an approximately equal number of healthy controls, enable an estimation of the sensitivity with a margin of error of 0.06 at the desired specificity of 0.95 or greater. The experiments were not randomized, and investigators were not blinded to allocation during experiments and outcome assessment.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

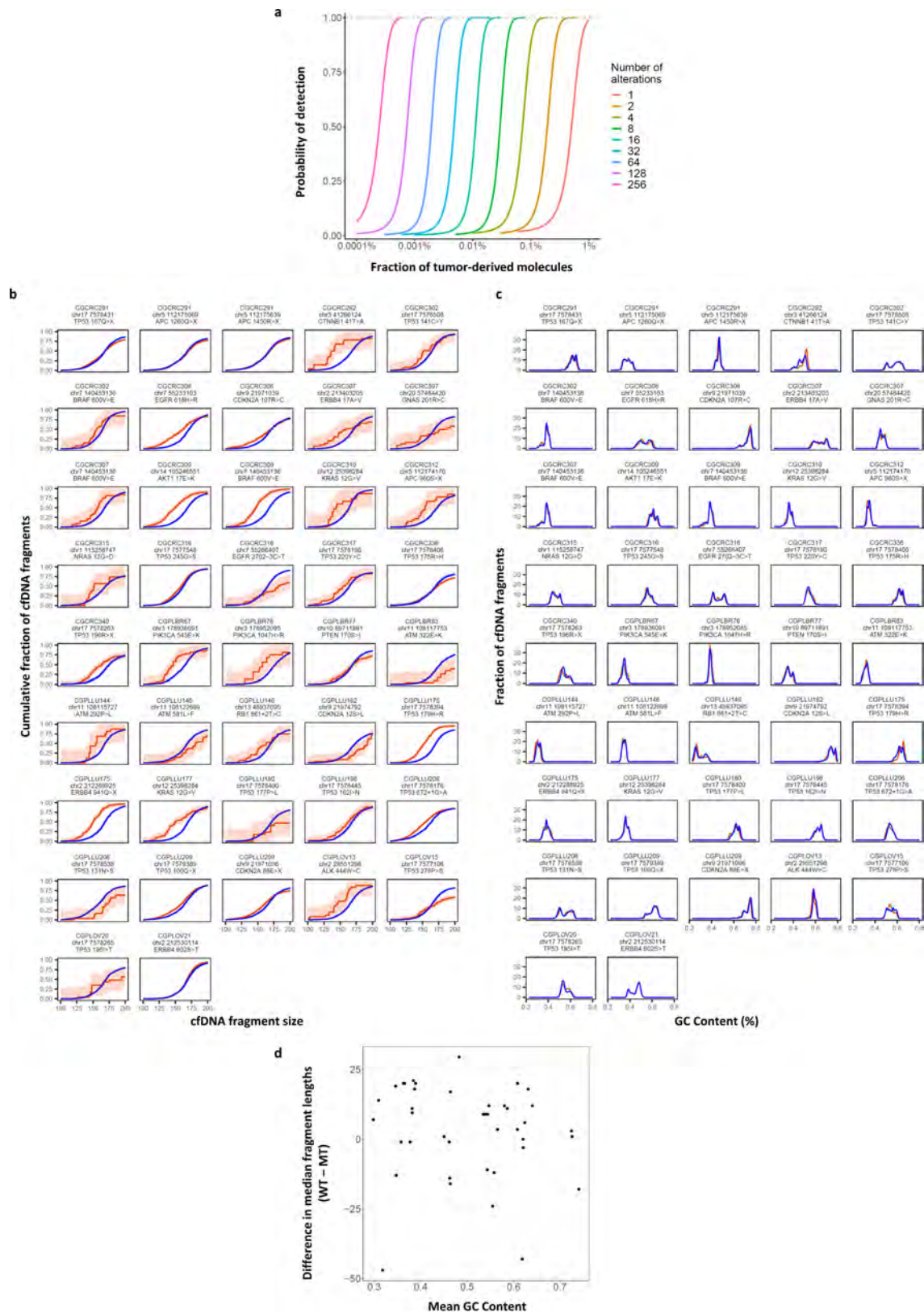
Data availability

Sequence data used in this study have been deposited at the database of Genotypes and Phenotypes (dbGaP, study ID 34536).

Code availability

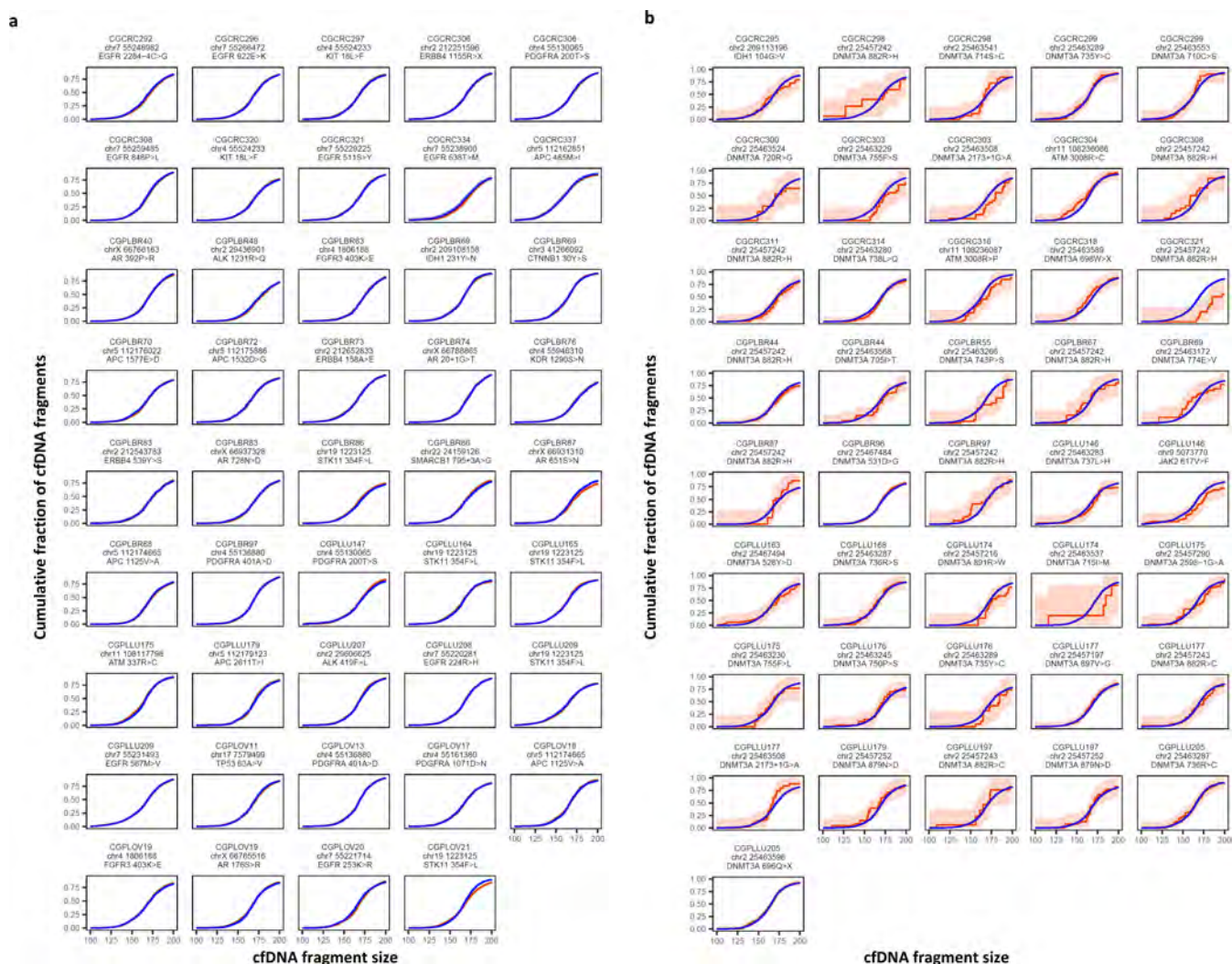
Code for analyses is available at http://github.com/Cancer-Genomics/delfi_scripts.

32. Fisher, S. et al. A scalable, fully automated process for construction of sequence-ready human exome targeted capture libraries. *Genome Biol.* **12**, R1 (2011).
33. Jones, S. et al. Personalized genomic analyses for cancer mutation discovery and interpretation. *Sci. Transl. Med.* **7**, 283ra53 (2015).
34. Benjamini, Y. & Speed, T. P. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res.* **40**, e72 (2012).
35. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
36. Friedman, J. H. Greedy function approximation: a gradient boosting machine. *Ann. Stat.* **29**, 1189–1232 (2001).
37. Friedman, J. H. Stochastic gradient boosting. *Comput. Stat. Data Anal.* **38**, 367–378 (2002).
38. Efron, B. & Tibshirani, R. Improvements on cross-validation: the 632+ bootstrap method. *J. Am. Stat. Assoc.* **92**, 548–560 (1997).
39. Zurbenko, I. G. *The Spectral Analysis of Time Series* (Elsevier, 1986).
40. Robin, X. et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* **12**, 77 (2011).



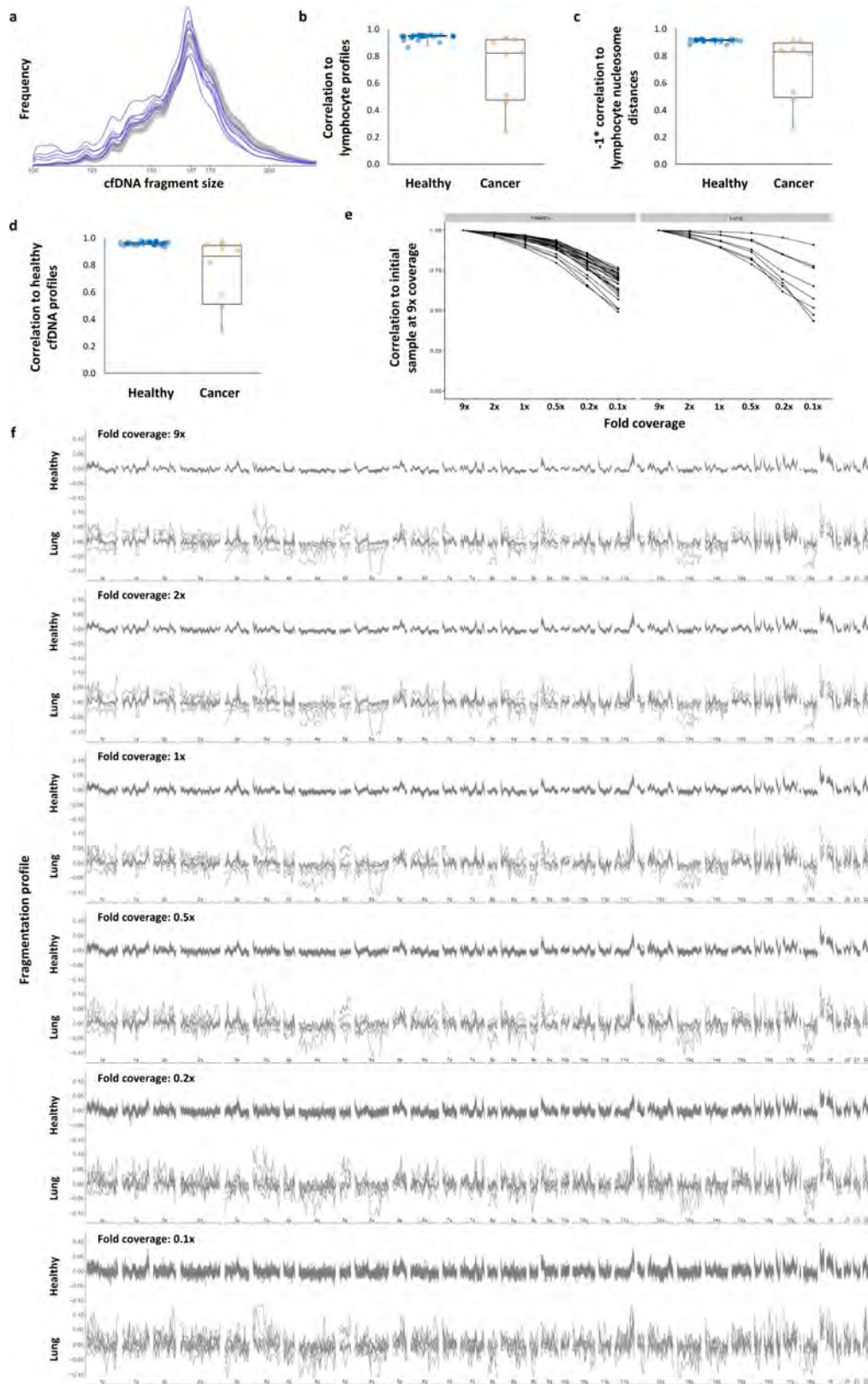
Extended Data Fig. 1 | Simulations of non-invasive cancer detection based on number of alterations analysed and tumour-derived cfDNA fragment distributions. **a**, Monte Carlo simulations were performed using different numbers of tumour-specific alterations to evaluate the probability of detecting cancer alterations in cfDNA at the indicated fraction of tumour-derived molecules. The simulations were performed assuming an average of 2,000 genome equivalents of cfDNA and the requirement of five or more observations of any alteration. These analyses indicate that increasing the number of tumour-specific alterations

improves the sensitivity of detection of circulating tumour DNA. **b**, Cumulative density functions of cfDNA fragment lengths of 42 loci containing tumour-specific alterations from 30 patients with breast, colorectal, lung, or ovarian cancer are shown with 95% confidence bands (orange). Lengths of mutant cfDNA fragments were significantly different in size from wild-type cfDNA fragments (blue) at these loci. **c**, GC content was similar for mutated and non-mutated fragments. **d**, GC content was not correlated to fragment length.



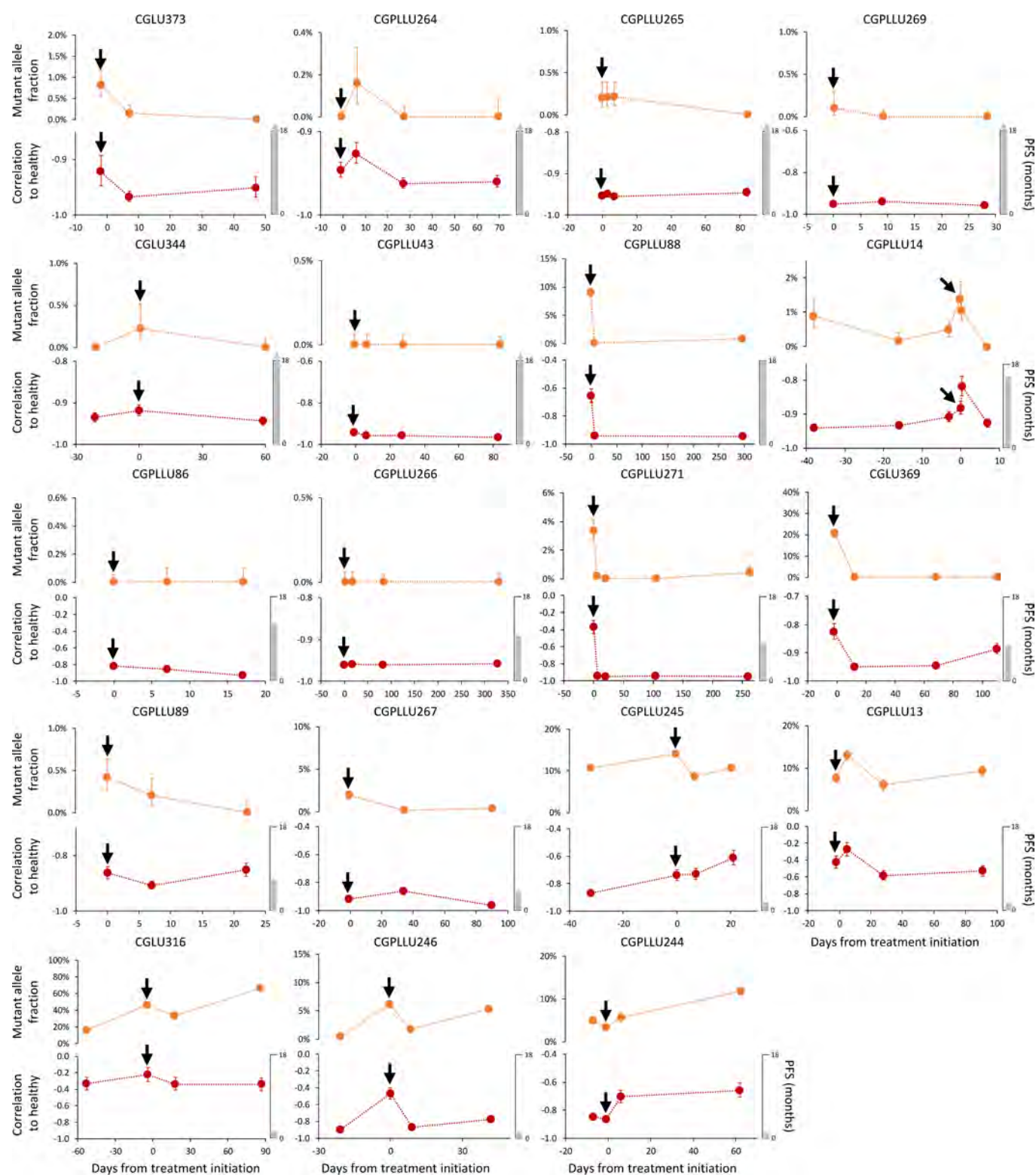
Extended Data Fig. 2 | Germline and haematopoietic cfDNA fragment distributions. **a**, Cumulative density functions of fragment lengths at 44 loci containing germline alterations (non-tumour derived) from 38 patients with breast, colorectal, lung or ovarian cancer are shown with 95% confidence bands. Fragments with germline mutations (orange) were comparable in length to wild-type cfDNA fragment lengths (blue).

b, Cumulative density functions of fragment lengths at 41 loci containing haematopoietic alterations (non-tumour derived) from 28 patients with breast, colorectal, lung or ovarian cancer are shown with 95% confidence bands. After correction for multiple testing, there were no significant differences ($\alpha = 0.05$) in the size distributions of mutated haematopoietic cfDNA fragments (orange) and wild-type cfDNA fragments (blue).



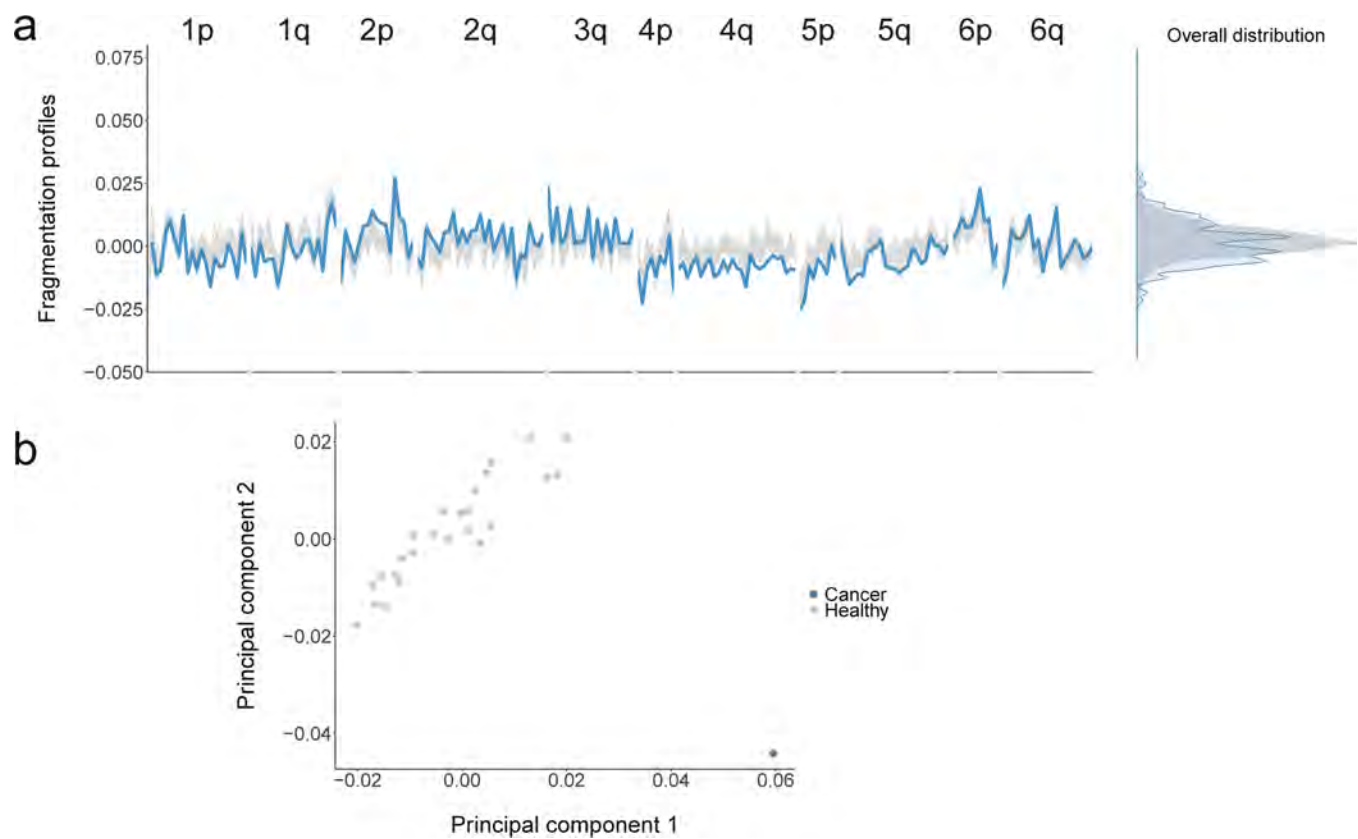
Extended Data Fig. 3 | cfDNA fragmentation in healthy individuals and patients with lung cancer. **a**, cfDNA fragment lengths are shown for healthy individuals ($n = 30$, grey) and patients with lung cancer ($n = 8$, blue). **b–d**, cfDNA fragmentation profiles from healthy individuals ($n = 30$) had high correlations, whereas patients with lung cancer ($n = 8$) had lower correlations to median fragmentation profiles of lymphocytes (**b**), lymphocyte nucleosome distances (**c**) and healthy cfDNA (**d**). Pearson correlations are shown with box plots depicting minimum, 25th percentile,

median, 75th percentile, and maximum values. **e**, High coverage (9x) WGS data were subsampled to 2x, 1x, 0.5x, 0.2x and 0.1x-fold coverage. Mean centred genome-wide fragmentation profiles in 5-Mb bins for 30 healthy individuals and 8 patients with lung cancer are depicted for each subsampled fold coverage with median profiles shown in blue. **f**, Pearson correlation of subsampled profiles to initial profile at 9x coverage for healthy individuals and patients with lung cancer.



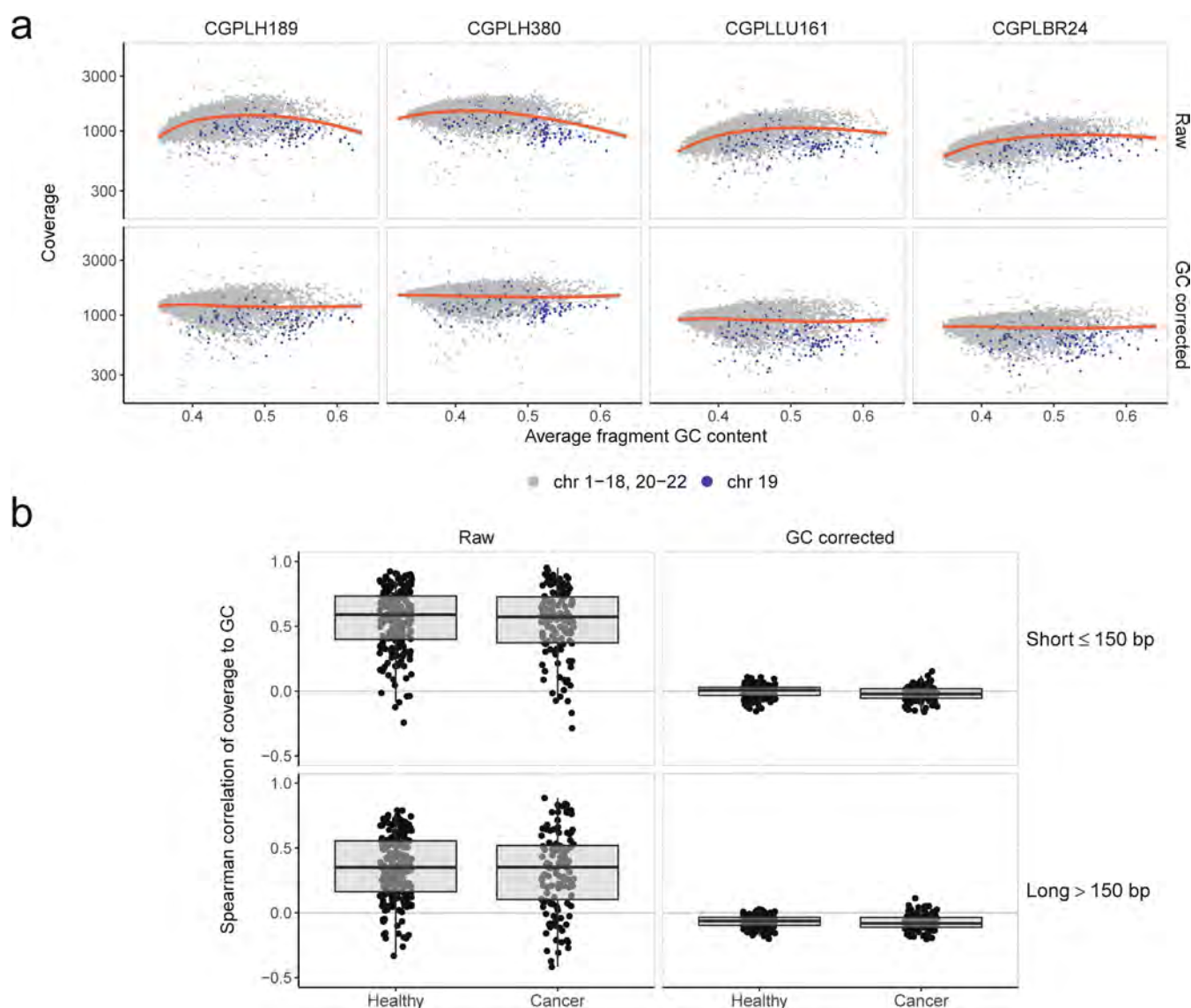
Extended Data Fig. 4 | cfDNA fragmentation profiles and sequence alterations during therapy. Detection and monitoring of cancer in serial blood draws from patients with non-small cell lung cancer ($n = 19$) undergoing treatment with targeted tyrosine kinase inhibitors (black arrows) was performed using targeted sequencing (top) as previously reported²⁹, and genome-wide fragmentation profiles (bottom). For each case, the vertical axis of the bottom panel displays -1 times the Pearson correlation of each sample to the median healthy cfDNA fragmentation profile. Error bars depict confidence intervals from binomial tests for mutant allele fractions, and confidence intervals calculated using Fisher

transformation for genome-wide fragmentation profiles. Although the approaches analyse different aspects of cfDNA (whole genome compared with specific alterations), the targeted sequencing and fragmentation profiles were similar for patients responding to therapy as well as those with stable or progressive disease. As fragmentation profiles reflect both genomic and epigenomic alterations (whereas mutant allele fractions only reflect individual mutations), mutant allele fractions alone may not reflect the absolute level of correlation of fragmentation profiles to healthy individuals.



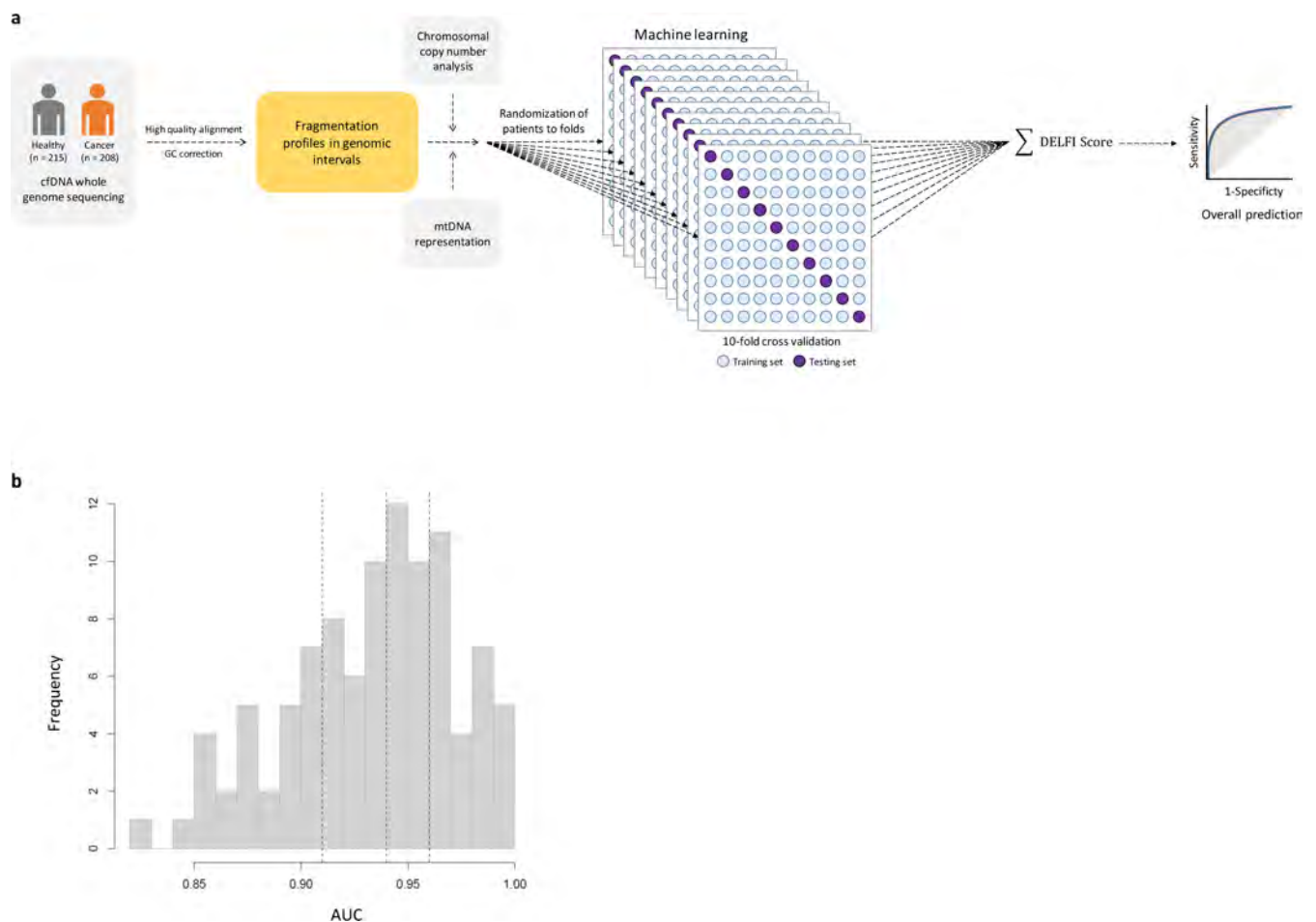
Extended Data Fig. 5 | Profiles of cfDNA fragment lengths in copy neutral regions in healthy individuals and one patient with colorectal cancer. **a**, The fragmentation profiles in 211 copy neutral windows in chromosomes 1–6 are shown for 25 randomly selected healthy individuals (grey). For a patient with colorectal cancer (CGCRC291) with an estimated mutant allele fraction of 20%, we diluted the cancer fragment length profile to an approximate 10% tumour contribution (blue). **a**, **b**,

Although the marginal densities of the fragment profiles for the healthy samples and patient with cancer show substantial overlap (**a**, right), the fragmentation profiles are different as can be seen through visualization of the fragmentation profiles (**a**, left) and by the separation of the patient with colorectal cancer from the healthy samples ($n = 25$) in a principal component analysis (**b**).



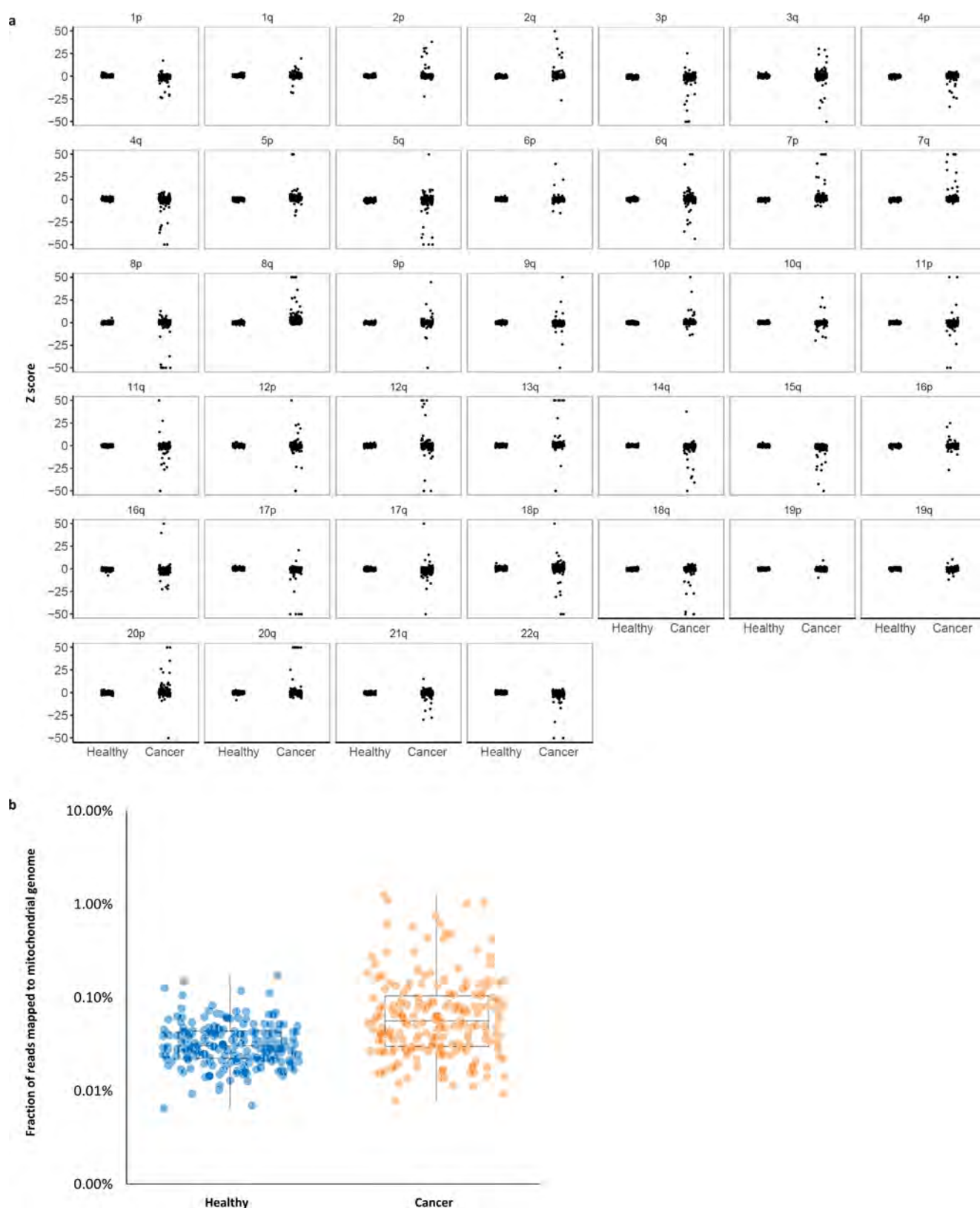
Extended Data Fig. 6 | Genome-wide GC correction of cfDNA fragments. To estimate and control for the effects of GC content on sequencing coverage, we calculated coverage in non-overlapping 100-kb genomic windows across the autosomes. For each window, we calculated the average GC of the aligned fragments. **a**, LOESS smoothing of raw coverage (top row) for two randomly selected healthy subjects (CGPLH189 and CGPLH380) and two patients with cancer (CGPLLU161 and CGPLBR24) with undetectable aneuploidy (PA score < 2.35). After subtracting the average coverage predicted by the LOESS model, the residuals were rescaled to the median autosomal coverage (bottom row). As fragment length may also result in coverage biases, we performed this

GC correction procedure separately for short (≤ 150 bp) and long (> 150 bp) fragments. Although the 100-kb bins on chromosome 19 (blue points) consistently have less coverage than predicted by the LOESS model, we did not implement a chromosome-specific correction as such an approach would remove the effects of chromosomal copy number on coverage. **b**, Overall, we found a limited correlation between short or long fragment coverage and GC content after correction among healthy individuals ($n = 211$, interquartile range: -0.03 – 0.03) and patients with cancer ($n = 128$, interquartile range: -0.06 – 0.02) with a PA score < 3 . Box plots depict 25th percentile, median, and 75th percentile values.



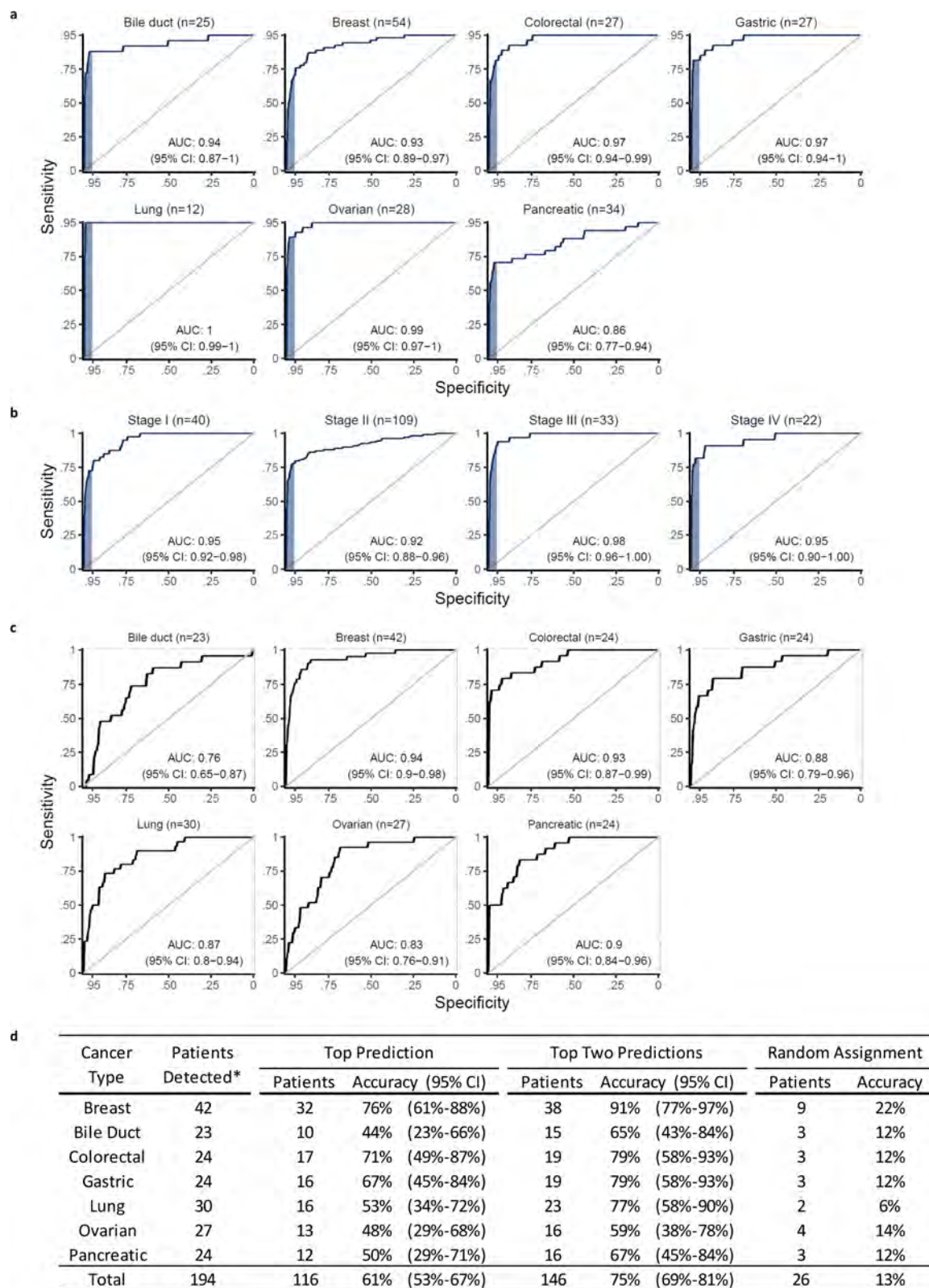
Extended Data Fig. 7 | Machine learning model. **a**, We used gradient tree boosting machine learning to examine whether cfDNA can be categorized as having characteristics of a patient with cancer or a healthy individual. The machine learning model included fragmentation size and coverage characteristics in windows throughout the genome, as well as chromosomal arm and mitochondrial DNA copy numbers. We used a tenfold cross-validation approach in which each sample is randomly assigned to a fold, and nine of the folds (90% of the data) are used for training and one fold (10% of the data) is used for testing. The prediction accuracy from a single cross-validation is an average over the ten possible combinations of test and training sets. As this prediction accuracy can

reflect bias from the initial randomization of patients, we repeat the entire procedure, including the randomization of patients to folds, ten times. For all cases, feature selection and model estimation were performed on training data and were validated on test data, and the test data were never used for feature selection. Ultimately, we obtained a DELFI score that could be used to classify individuals as likely to be healthy or having cancer. **b**, Distribution of AUCs across the repeated tenfold cross-validation. The 25th, 50th and 75th percentiles of the 100 AUCs for the cohort of 215 healthy individuals and 208 patients with cancer are indicated by dashed lines.



Extended Data Fig. 8 | Whole-genome analyses of chromosomal arm copy number changes and mitochondrial genome representation. **a**, Z-scores for each autosome arm are depicted for healthy individuals ($n = 215$) and patients with cancer ($n = 208$). The vertical axis depicts normal copy at zero with positive and negative values indicating arm

gains and losses, respectively. Z-scores greater than 50 or less than -50 are thresholded at the indicated values. **b**, The fraction of reads mapping to the mitochondrial genome is depicted for healthy individuals ($n = 215$) and patients with cancer ($n = 208$). Box plots depict the minimum, 25th percentile, median, 75th percentile, and maximum values.



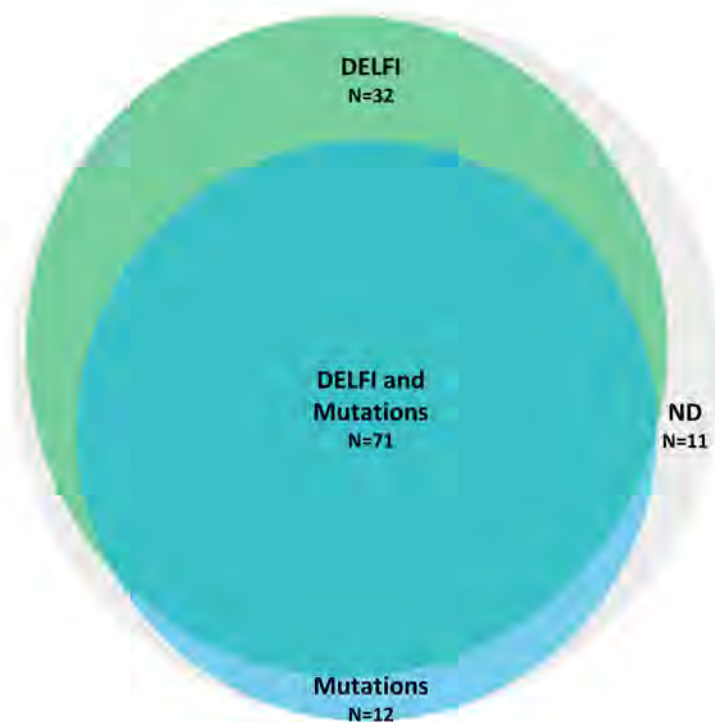
*Patients detected are based on DELFI detection at 90% specificity. Lung cohort includes additional lung cancer patients with prior therapy.

Extended Data Fig. 9 | DELFI detection of cancer and tissue of origin prediction. **a**, Analyses of individual cancer types using DELFI had AUCs ranging from 0.86 to >0.99. **b**, Receiver operator characteristics for detection of cancer using cfDNA fragmentation profiles and other genome-wide features in a machine learning approach are depicted for a cohort of 215 healthy individuals and each stage of 208 patients with cancer with $\geq 95\%$ specificity shaded in blue. **c**, Receiver operator

characteristics for DELFI tissue prediction of bile duct, breast, colorectal, gastric, lung, ovarian or pancreatic cancer are depicted. To increase sample sizes within cancer type classes, we included cases detected with a 90% specificity, and the lung cancer cohort was supplemented with the addition of baseline cfDNA data from 18 patients with lung cancer with prior treatment³⁶. **d**, DELFI tissue of origin prediction.

Detection approach*		Patients analyzed	Patients detected	Fraction of patients detected	95% CI
Stage	DELFI	126	103	82%	74%-88%
	Mutations	126	83	66%	57%-74%
	DELFI and Mutations	126	115	91%	85%-96%
	I	32	27	84%	67%-95%
	II	52	48	92%	81%-98%
	III	25	23	92%	74%-99%
	IV	16	16	100%	79%-100%

*Cancer detection using DELFI, sequence mutations, and the combination of DELFI and mutations was performed at specificities of 98%, >99%, and 98%, respectively. Per stage sensitivities are included for all cases except for one patient with stage X.



Extended Data Fig. 10 | Detection of cancer using DELFI and mutation-based cfDNA approaches. DELFI (green) and targeted sequencing¹⁰ for mutation identification (blue) were performed independently in a cohort of 126 patients with breast, bile duct, colorectal, gastric, lung or ovarian

cancer. The number of individuals detected by each approach and in combination are indicated for DELFI detection with a specificity of 98%, targeted sequencing specificity at >99%, and a combined specificity of 98%. ND, not detected.

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a Confirmed

- ☐ ☒ The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- ☐ ☒ An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- ☐ ☒ The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- ☐ ☒ A description of all covariates tested
- ☐ ☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- ☐ ☒ A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- ☐ ☒ For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
Give P values as exact values whenever suitable.
- ☒ ☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- ☒ ☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- ☐ ☒ Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated
- ☐ ☒ Clearly defined error bars
State explicitly what error bars represent (e.g. SD, SE, CI)

Our web collection on [statistics for biologists](#) may be useful.

Software and code

Policy information about [availability of computer code](#)

Data collection

Primary processing of whole genome NGS data for cfDNA samples was performed using Illumina CASAVA (version 1.8.2) with alignment using ELAND. Re-alignment of mitochondrial mapped reads was done using Bowtie-2 (version 2.3.4) in end-to-end mode. Sequence reads in the analysis of targeted data were aligned using NovoAlign (version 3.02.12) and variant calling was performed using VariantDx.

Data analysis

All analyses were performed using R (version 3.4.3). Binning of the human genome was done using the R package GenomicRanges (version 1.30.3). Prediction was implemented using R package gbm (version 2.1.4), caret (version 6.0.79) and pROC (version 1.13). Bam file processing and analyses of targeted data were performed using Rsamtools (version 1.30.0) and GenomicAlignments (version 1.14.2).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Sequence data utilized in this study have been deposited at the database of Genotypes and Phenotypes (dbGaP, study ID 34536). Code for analyses is available at http://github.com/Cancer-Genomics/delfi_scripts.

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/authors/policies/ReportingSummary-flat.pdf](https://www.nature.com/authors/policies/ReportingSummary-flat.pdf)

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Assuming the prevalence of undiagnosed cancer cases in this population is high (1 or 2 cases per 100 healthy), a genomic assay with a specificity of 0.95 and sensitivity of 0.8 would have useful operating characteristics (positive predictive value of 0.25 and negative predictive value near 1). An analysis of more than 200 cancer patients and an approximately equal number of healthy controls would provide an estimation of the sensitivity with a margin of error of 0.06 at the desired specificity of 0.95.
Data exclusions	Plasma samples that were not collected according the described protocol were not used in the analysis.
Replication	We successfully performed internal replication in our study, including replication of the initial observations of our pilot study in a larger analysis of healthy individuals and patients with cancer.
Randomization	Not applicable
Blinding	Not applicable

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Unique biological materials
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input type="checkbox"/>	<input checked="" type="checkbox"/> Human research participants

Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

Human research participants

Policy information about [studies involving human research participants](#)

Population characteristics	Plasma samples from healthy individuals and plasma and tissue samples from patients with breast, lung, ovarian, colorectal, bile duct, and gastric cancers were obtained from ILSBio/Bioreclamation, Aarhus University, Herlev Hospital of the University of Copenhagen, Hvidovre Hospital, the University Medical Center of the University of Utrecht, the Academic Medical Center of the University of Amsterdam, the Netherlands Cancer Institute, and the University of California, San Diego. All samples were obtained under Institutional Review Board approved protocols with informed consent for research use at participating institutions. Plasma samples from healthy individuals were obtained at the time of routine screening, including for
----------------------------	--

Recruitment

colonoscopies or Pap smears. Individuals were considered healthy if they had no previous history of cancer and negative screening results.

Participants were recruited through screening trials, observational trials, or through formal biospecimen collection at University center hospitals. Potential self-selection bias or other biases were not identified.

De novo protein design by citizen scientists

Brian Koepnick^{1,2}, Jeff Flatten³, Tamir Husain³, Alex Ford^{1,2}, Daniel-Adriano Silva^{1,2}, Matthew J. Bick^{1,2}, Aaron Bauer³, Gaohua Liu^{4,5}, Yojiro Ishida⁶, Alexander Boykov¹¹, Roger D. Estep¹¹, Susan Kleinfelter¹¹, Toke Nørgård-Solano¹¹, Linda Wei¹¹, Foldit Players¹⁰, Gaetano T. Montelione^{4,6}, Frank DiMaio^{1,2}, Zoran Popović³, Firas Khatib⁷, Seth Cooper⁸ & David Baker^{1,2,9*}

Online citizen science projects such as GalaxyZoo¹, Eyewire² and Phylo³ have proven very successful for data collection, annotation and processing, but for the most part have harnessed human pattern-recognition skills rather than human creativity. An exception is the game EteRNA⁴, in which game players learn to build new RNA structures by exploring the discrete two-dimensional space of Watson-Crick base pairing possibilities. Building new proteins, however, is a more challenging task to present in a game, as both the representation and evaluation of a protein structure are intrinsically three-dimensional. We posed the challenge of de novo protein design in the online protein-folding game Foldit⁵. Players were presented with a fully extended peptide chain and challenged to craft a folded protein structure and an amino acid sequence encoding that structure. After many iterations of player design, analysis of the top-scoring solutions and subsequent game improvement, Foldit players can now—starting from an extended polypeptide chain—generate a diversity of protein structures and sequences that encode them in silico. One hundred forty-six Foldit player designs with sequences unrelated to naturally occurring proteins were encoded in synthetic genes; 56 were found to be expressed and soluble in *Escherichia coli*, and to adopt stable monomeric folded structures in solution. The diversity of these structures is unprecedented in de novo protein design, representing 20 different folds—including a new fold not observed in natural proteins. High-resolution structures were determined for four of the designs, and are nearly identical to the player models. This work makes explicit the considerable implicit knowledge that contributes to success in de novo protein design, and shows that citizen scientists can discover creative new solutions to outstanding scientific challenges such as the protein design problem.

The principle underlying de novo protein design is that proteins fold to their lowest free-energy state⁶; hence, designing a new protein structure requires finding an amino acid sequence with its lowest energy state in the prescribed structure. In practice, this challenge can be divided into two subproblems: first, crafting a protein backbone that is designable (that is, that could be the lowest energy state of some sequence); and second, finding a sequence with its lowest energy state in the crafted structure. One of the challenges of protein design is the exponentially increasing number of conformations available to a polypeptide chain, which is huge even for a modestly sized protein of 60–100 residues. Thus, the first subproblem of crafting a plausible backbone is extremely open-ended, and the second subproblem is difficult because it is not tractable to explicitly check that a designed sequence has lower energy in the crafted structure than in any other structure. There has been considerable progress in de novo protein design in recent years^{7–10}, but it is unclear whether all of the contributions to this success have been made explicit in the protocols used to design proteins, and how much implicit knowledge resides in the expertise of

the designers. Disentangling the role of expert knowledge is particularly difficult for the extremely open-ended challenge posed by the first subproblem (that is, crafting a plausible backbone), for which there are a practically unlimited number of solutions. Because full computer enumeration of backbones is not possible, there is considerable room for human creativity and intuition in generating and designing new protein structures.

To investigate how crowd-based creativity could contribute to solving the de novo protein design problem, we incorporated de novo design tools into the protein-folding game Foldit. Foldit is a free online computer game developed to crowdsource problems in protein modelling, and provides full control over the three-dimensional structure of a protein model⁵ (Fig. 1). Players compete to build a model with the lowest free energy, as calculated using the Rosetta energy function¹¹. In the past, Foldit has been primarily applied to protein structure prediction problems, in which players are presented with an unstructured amino acid sequence and challenged to determine its native conformation^{5,12}. In one case, Foldit players redesigned a loop region of an already folded structure¹³, but the de novo design of an entire protein is a far more expansive challenge.

We repeatedly challenged Foldit players to design stably folded proteins from scratch, and iteratively improved the game on the basis of their results. In each challenge, players were provided with a polyisoleucine backbone in a fully extended conformation (60–100 residues in length) and were given 7 days to fold the backbone into a compact structure and identify a sequence specifying this backbone. Initially, most top-scoring (low-energy) Foldit player designs were highly extended, lacked a solvent-inaccessible core and were composed entirely of polar residues (Extended Data Fig. 1); such extended, fully α -helical structures have more favourable hydrogen bonding, electrostatic and local torsional energies than collapsed structures, which must contort to create a buried core. Whereas polylysine and other extended polar sequences resembling these initial Foldit solutions are often α -helical in solution^{14,15}, the lack of long-range interactions precludes specific folding into a single stable structure¹⁶. This highlights a limitation of using absolute energy as an optimization criterion for protein design: a low-energy design does not guarantee structural specificity, which arises only if all other alternative conformations have higher energy. To favour the design of globular solvent-excluding protein folds, with sequences that uniquely encode them, we introduced three supplementary design rules into Foldit: a ‘core exists’ rule that requires a minimum proportion of residues (for example, 30%) to be solvent-inaccessible in the designed structure; a ‘secondary structure design’ rule that prohibits glycine and alanine in all secondary structure elements; and a ‘residue interaction energy’ rule to penalize large residues that make insufficient intramolecular interactions in the designed structure. With the addition of these rules to Foldit, subsequent top-scoring designs from Foldit players were compact globular proteins.

¹Department of Biochemistry, University of Washington, Seattle, WA, USA. ²Institute for Protein Design, University of Washington, Seattle, WA, USA. ³Center for Game Science, Paul G. Allen School of Computer Science and Engineering, University of Washington, Seattle, WA, USA. ⁴Department of Molecular Biology and Biochemistry, Rutgers University The State University of New Jersey, Piscataway, NJ, USA. ⁵Nexomics Biosciences, Bordentown, NJ, USA. ⁶Department of Biochemistry, Robert Wood Johnson Medical School, Rutgers The State University of New Jersey, Piscataway, NJ, USA. ⁷Department of Computer and Information Science, University of Massachusetts Dartmouth, Dartmouth, MA, USA. ⁸Khoury College of Computer Sciences, Northeastern University, Boston, MA, USA. ⁹Howard Hughes Medical Institute, University of Washington, Seattle, WA, USA. ¹⁰A list of participants appears in the Supplementary Information. ¹¹Unaffiliated: Alexander Boykov, Roger D. Estep, Susan Kleinfelter, Toke Nørgård-Solano, Linda Wei. *e-mail: dabaker@uw.edu

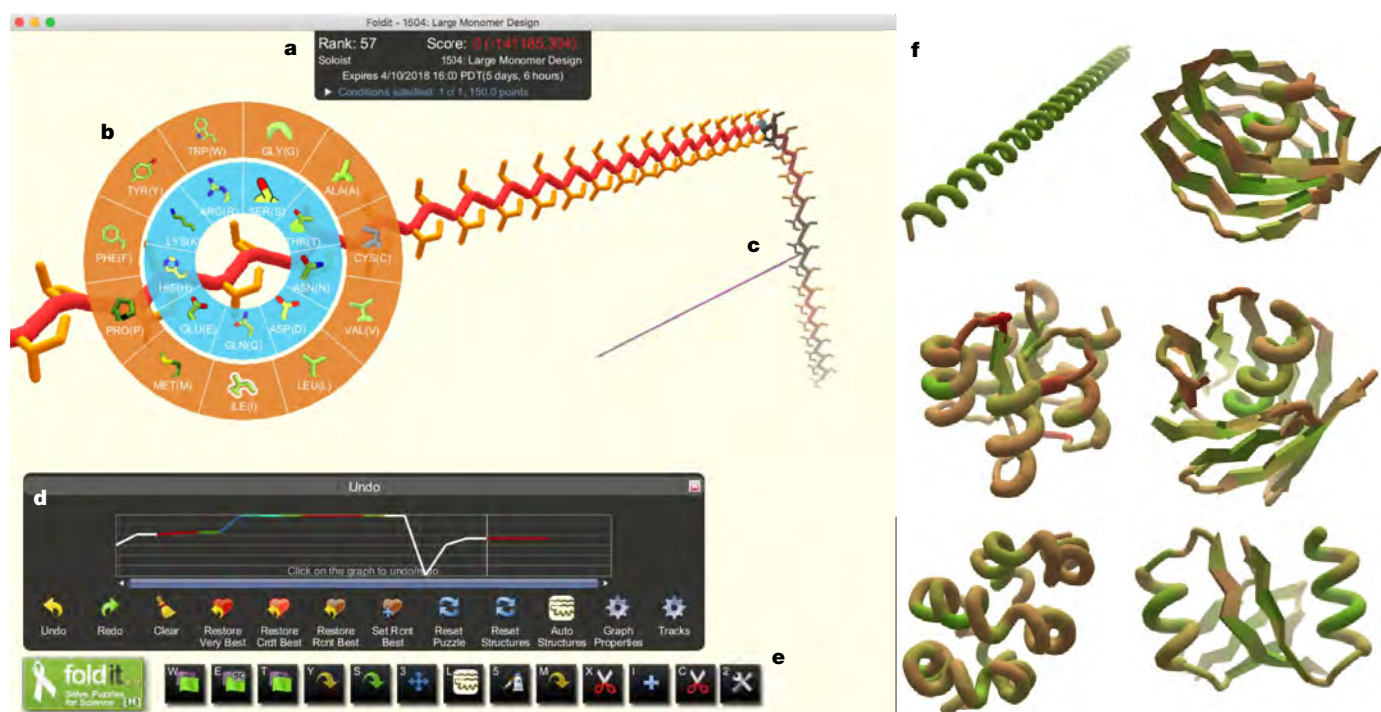


Fig. 1 | The Foldit user interface. **a**, The Foldit score is the Rosetta energy with a negative multiplier, so that better models yield higher scores.

b, The design palette allows players to change the identity of the amino-acid residue at any position of the model. **c**, The 'pull' tool allows players to manipulate the 3D structure of the model. **d**, The 'undo' graph tracks the score as a model is developed, and allows players to backtrack and load previous versions of a model. **e**, Additional Foldit tools (from left to

right): full-structure minimization, sidechain minimization, backbone minimization, auto-design sidechains, repack sidechains, translate or rotate model, secondary structure assignment, idealize secondary structure, manually design sidechains, delete residues, insert residues, insert cutpoint and idealize peptide bond geometry. **f**, Foldit players explore diverse structures that have no sequence or structural homology to natural proteins.

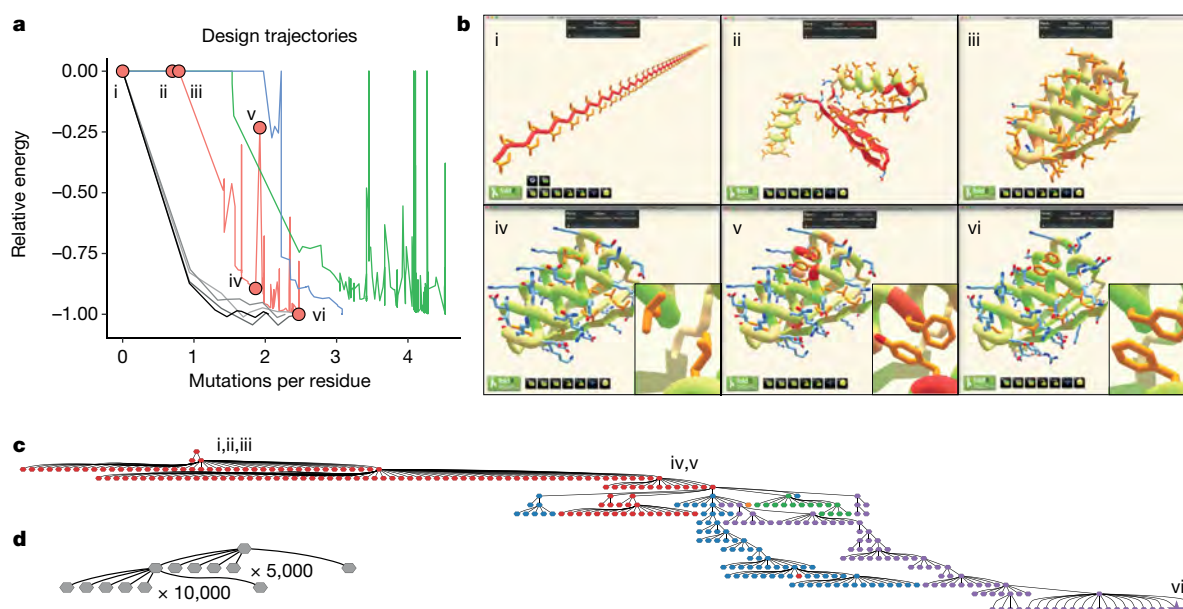


Fig. 2 | Comparison of Foldit player and automated design-sampling strategies. **a**, Single trajectories (ignoring abandoned branches) for three Foldit player-designed proteins in red (Foldit1), blue (Peak6) and green (Ferredox-Diesel); and design trajectories for four Rosetta-designed proteins in grey. The y axis is the Rosetta energy rescaled so that the final design has a value of -1.00 , and positive energies are shown as zero. Foldit players are willing to undergo large increases in energy to explore new regions; by contrast, the Rosetta protocol has a limited ability to escape local energy minima. Red circles correspond to structures shown in **b**. **b**, Snapshots from the design trajectory of Foldit1: (i) the initial extended chain of polyisoleucine; (ii) development of secondary structure; (iii) development of folded tertiary

structure; (iv) sequence design of folded structure, with inset showing favourable packing at positions 13 and 45; (v) high-energy intermediate design, with inset showing redesign at positions 13 and 45, which results in steric clashes with the protein backbone; (vi) the final refined design, with inset showing renewed favourable interactions at positions 13 and 45. **c**, The design strategy for Foldit1 represented as a graph, showing all branch points where multiple design trajectories were spawned from a single intermediate. The final design was reached after 17 branch points. Node colours correspond to five different cooperating Foldit players, and the final design is marked with a star. **d**, Similar representation of a Rosetta design trajectory—there are only two branch points.

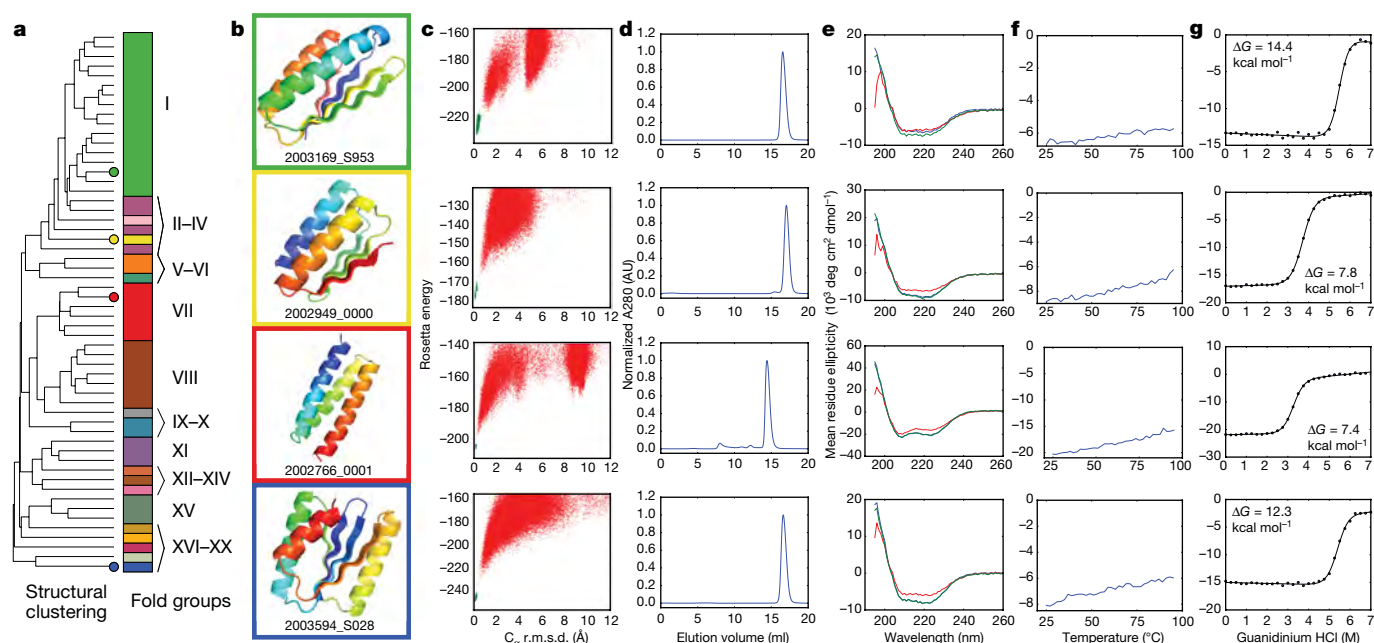


Fig. 3 | Structural characterization of Foldit player-designed proteins. **a**, Dendrogram showing all 56 folded Foldit player designs clustered by structural similarity (TM-align²⁶), with coloured circles highlighting the four designs characterized in **b–g**. The stacked bars show the 20 different folds among the clustered designs (Extended Data Fig. 5). Fold XX (see design 2003594_S028) is a new fold, previously unobserved in natural proteins. **b–g**, Cartoon depiction of four select Foldit designs (**b**); the graphs in **c–g** correspond to these four structures. **c**, Rosetta@home ab initio calculations show that the sequence for each design has an energy landscape that is strongly funnelled towards the design structure. *y* axis, Rosetta energy; *x* axis, C_{α} r.m.s.d. to the designed structure; points represent lowest-energy structures sampled starting from an extended chain (red points) and starting from the Foldit design model

We obtained custom synthetic genes encoding 12 player designs for which structure prediction calculations converged on the player-designed conformation¹⁷. The sequences of these proteins have no homology to any known protein (Supplementary Table 1). The de novo designs were expressed in *E. coli* and purified by metal-affinity and size-exclusion chromatography. Analysis by chromatography and circular dichroism indicated that 6 of the 12 designs were monomeric and folded in solution, with helical secondary structure consistent with the players' models (Supplementary Fig. 1). All of the experimentally tested proteins described in this paper are entirely the work of Foldit players.

During gameplay, the Foldit application uploads the player's latest model to the Foldit server every 2–5 minutes; from these snapshots we can reconstruct the process by which a Foldit player develops a protein design (Fig. 2). Foldit players use more-varied and complex exploration strategies than standard Rosetta automated design protocols, and frequently revert to a previous iteration of their model to explore an alternative path, resulting in a highly branched search tree. A typical automated design protocol, by contrast, includes only two branch points¹⁸. In addition, Foldit players regularly sample much higher energy states than the automated protocol, which has only a limited ability to escape local energy minima.

Encouraged by the success of Foldit players in designing stable proteins from scratch, we made additions to the game to encourage players to explore more-diverse protein structures. Up until this point, all top-scoring Foldit designs had consisted of either three or four α -helices connected by minimal loops. Indeed, Foldit players had determined that designs with β -sheets did not score as well as α -helical bundles (Extended Data Fig. 2), and competitive players had abandoned any attempt to design more varied folds. This is an interesting

(green points). **d**, Size-exclusion chromatography traces (absorbance at 280 nm) show that designs are monomeric in solution. **e**, Circular dichroism spectra indicate that the designs adopt the expected secondary structure content in solution at 25 °C (blue trace), when heated to 95 °C (red trace) and when cooled again to 25 °C (green trace). **f**, Circular dichroism mean residue ellipticity at 220 nm as temperature is increased from 25 °C to 95 °C; the designs do not denature with increasing temperature. **g**, Cooperative unfolding during titration with guanidinium hydrochloride. Blue circles show circular dichroism mean residue ellipticity at 220 nm with increasing concentration of denaturant, and the black curve shows a two-state unfolding model fit to the data. Free energy of unfolding (ΔG_{unf}) was determined by linear extrapolation using the fit model parameters²⁷.

parallel to protein design by practicing scientists, which has also focused much more on helical bundles than on other classes of protein folds^{19–22}. To encourage the design of a wider variety of folds, we introduced a 'secondary structure' rule, stipulating that no more than 50% of residues may form α -helices. Foldit players responded by designing a multitude of mixed α/β -proteins, which were indistinguishable from expert designs on visual inspection. However, structure prediction calculations for these α/β design sequences showed poor sampling close to the target design structure, which suggests that the designed sequences did not strongly encode their local structures¹⁷. Further analysis showed that these player designs contained many residues with locally strained backbone conformations (backbone ϕ and ψ torsions in unfavoured regions of the Ramachandran plot^{23,24}). That such designs had very low energies revealed a problem in the Rosetta energy function at the time: because Rosetta users typically sampled backbones starting from fragments of native proteins, unfavourable local conformations were rarely encountered—therefore, it had not been discovered that the energies associated with local-backbone strain were being underestimated. We addressed this flaw in the Rosetta model by increasing the steepness of the energetic penalties associated with strained local-backbone geometry; this is now implemented in the latest Rosetta energy function¹¹. We also added to Foldit an 'ideal loops' rule that restricted players to a set of 19 unstrained reverse-turn conformations⁷, and incorporated new tools to aid generation of unstrained backbones: a fragment lookup-based loop-closure tool, an interactive Ramachandran map and a protein blueprint scheme for drag-and-drop assembly of secondary structure elements and common loop conformations (Extended Data Fig. 3). Together, these upgrades brought about a marked improvement in the local-backbone quality of Foldit player-designed proteins (Extended Data Fig. 4).

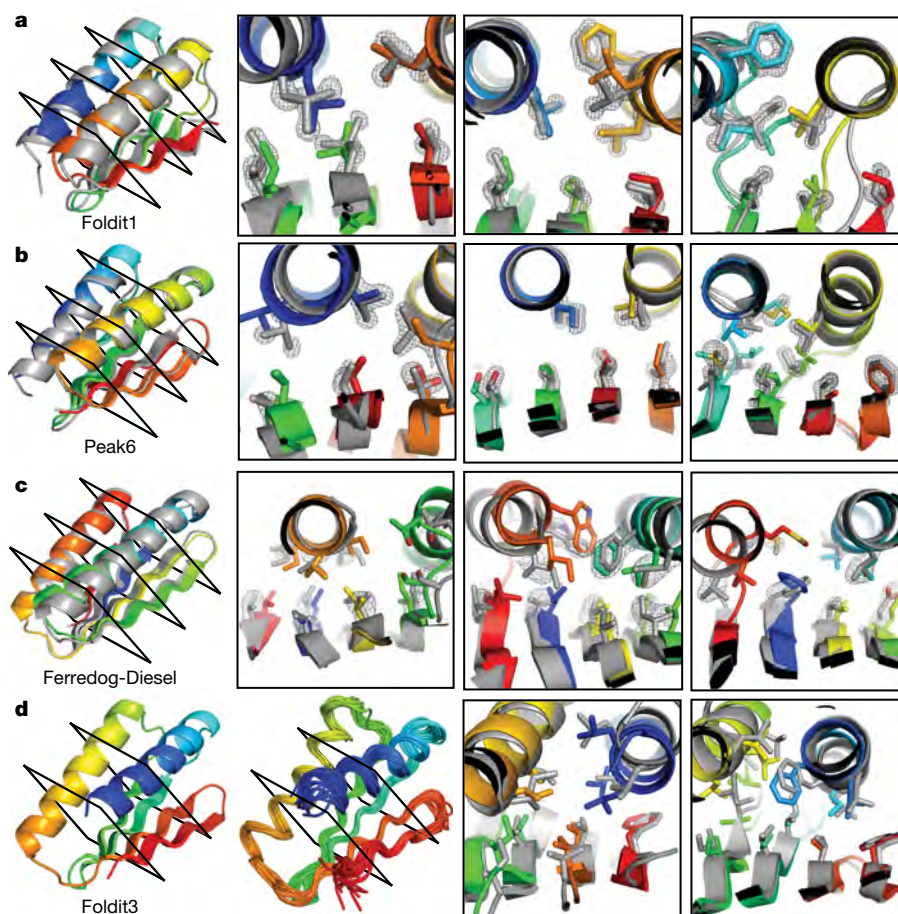


Fig. 4 | High-resolution structures of Foldit player-designed proteins. **a**, The Foldit1 design (fold V in Fig. 3: three β -strands with sheet order 1–2–3) model backbone (rainbow) aligns to the crystal structure (grey) with C_{α} r.m.s.d. of 1.1 Å. **b**, The Peak6 design (fold III: four strands, sheet order 1–2–4–3) model backbone aligns to the crystal structure with C_{α} r.m.s.d. of 0.9 Å. **c**, The Ferredog-Diesel design (fold I: four strands, sheet order 4–1–3–2) model backbone aligns to the crystal structure with C_{α}

r.m.s.d. of 1.7 Å. Cross sections show core-residue sidechains, with the composite omit $2mF_o - DF_c$ map contoured at 2.0σ (**a**, **b**) or 1.0σ (**c**).

d, The Foldit3 design model (fold XVII: four strands, sheet order 2–1–3–4) and NMR ensemble. The design model aligns to the representative (medoid) NMR model with a C_{α} r.m.s.d. of 1.1 Å. Cross sections compare core sidechains in the design model (rainbow) and representative NMR model (grey).

The importance of reducing local-backbone strain was borne out in experimental characterization. Before the backbone modelling improvements described in the previous paragraph, only 4 of 37 Foldit α/β -designs tested (11%) were monomeric and structured in solution. Following the backbone modelling additions, 46 of 97 (47%) were monomeric and exhibited the expected secondary structure in solution. Most showed exceptional stability in thermal and chemical denaturation experiments, with some free energies of unfolding (ΔG_{unf}) exceeding 20 kcal mol⁻¹; indeed, 32 designed proteins remained completely folded at 95 °C (Fig. 3, Supplementary Fig. 1). This success rate surpasses that in previous reports of designed α/β -proteins^{7,12}.

Overall, the 56 successful Foldit designs are diverse in structure, representing 20 different protein folds (Fig. 3, Extended Data Fig. 5), one of which is a new fold that is previously unobserved in natural proteins. The success of Foldit designs is not attributed to just one or two exceptional Foldit players, but is shared broadly by the Foldit community (Supplementary Table 1). The 56 successful designs were created by 36 different Foldit players (the most prolific player created 10 successful designs); 19 designs were created collaboratively by at least 2 cooperating players; and 5 successful designs were not top-scoring, but were nevertheless flagged by players as personal favourites. Foldit players lack formal expertise in protein modelling (Extended Data Fig. 6, Supplementary Notes), but knowledge and intuition gained from playing protein structure prediction puzzles in Foldit translated to success in de novo protein design (Extended Data Fig. 7).

We succeeded in solving high-resolution structures of four Foldit player-designed proteins. X-ray crystal structures of three designed proteins (named by their designers Foldit1, Peak6 and Ferredog-Diesel) closely match the designed conformations, with C_{α} root mean square deviations (r.m.s.d.) of 1.1, 0.9 and 1.7 Å, respectively (Fig. 4). Well-resolved electron density in the protein core of Foldit1 and Peak6 shows that most sidechains adopt the intended rotamers and preserve the designed packing interactions. The electron density of Ferredog-Diesel is less clear, but the protein backbone adopts the designed fold, and many core sidechains appear to pack as intended. The solution nuclear magnetic resonance (NMR) structure of a fourth design, Foldit3, also closely matches the design conformation, with a C_{α} r.m.s.d. of 1.1 Å between the design model and the medoid conformer²⁵ of the ensemble.

From these results, we can draw several general conclusions about scientific models, citizen science and the interplay between the two. First, a scientific model that holds within the domain space considered by practicing scientists may not hold outside of this domain. This is most vividly illustrated by the highly extended structures generated by Foldit players in their first de novo design efforts, and later by the structures with strained local geometry not previously sampled by Rosetta users. Second, for citizen scientists to make essential and creative scientific contributions through online gaming, the scoring function of the game must be an accurate representation of the science. In our initial iterations, Foldit did not present to players a sufficiently accurate and general model to allow them to robustly design new proteins,

even though the underlying Rosetta software had been used for protein design by practicing scientists. Third and most importantly, citizen science offers a powerful way to systematically improve a scientific model through iterations of model trial and model improvement. Human game players are exceptionally capable at finding and exploiting unanticipated solutions that are otherwise unexplored by experienced scientists, whose focus is not on getting a high score, but rather on solving their specific scientific problem.

We have demonstrated that non-expert citizen scientists, playing the online computer game Foldit, can accurately design completely new protein structures from scratch. Locally, players' solutions are physically plausible and resemble natural proteins, but globally, they are creative and diverse. Proteins designed by citizen-scientist Foldit players are by no measure inferior to those of expert protein designers: they fold accurately to the intended conformation, show exceptional folding stability and span a wide diversity of structures. This result is all the more impressive given that *de novo* protein design was an almost completely unsolved problem just a few years ago, and the diversity in protein folds spanned by the successful Foldit players' models considerably exceeds that in any previous protein design report, to our knowledge. The sustained success of Foldit players over a wide diversity of protein folds highlights the power of human creativity when guided by scientific understanding presented in a readily comprehensible form.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-019-1274-4>.

Received: 19 October 2018; Accepted: 14 May 2019;

Published online 5 June 2019.

- Lintott, C. J. et al. Galaxy Zoo: morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey. *Mon. Not. R. Astron. Soc.* **389**, 1179–1189 (2008).
- Kim, J. S. et al. Space-time wiring specificity supports direction selectivity in the retina. *Nature* **509**, 331–336 (2014).
- Kawrykow, A. et al. Phylo: a citizen science approach for improving multiple sequence alignment. *PLoS ONE* **7**, e31362 (2012).
- Lee, J. et al. RNA design rules from a massive open laboratory. *Proc. Natl Acad. Sci. USA* **111**, 2122–2127 (2014).
- Cooper, S. et al. Predicting protein structures with a multiplayer online game. *Nature* **466**, 756–760 (2010).
- Epstein, C. J., Goldberger, R. F. & Anfinsen, C. B. The genetic control of tertiary protein structure: studies with model systems. *Cold Spring Harb. Symp. Quant. Biol.* **28**, 439–449 (1963).
- Lin, Y.-R. et al. Control over overall shape and size in *de novo* designed proteins. *Proc. Natl Acad. Sci. USA* **112**, E5478–E5485 (2015).
- Huang, P.-S., Boyken, S. E. & Baker, D. The coming of age of *de novo* protein design. *Nature* **537**, 320–327 (2016).
- Marcos, E. et al. Principles for designing proteins with cavities formed by curved β sheets. *Science* **355**, 201–206 (2017).
- Dou, J. et al. *De novo* design of a fluorescence-activating β -barrel. *Nature* **561**, 485–491 (2018).
- Alford, R. F. et al. The Rosetta all-atom energy function for macromolecular modeling and design. *J. Chem. Theory Comput.* **13**, 3031–3048 (2017).
- Khatib, F. et al. Crystal structure of a monomeric retroviral protease solved by protein folding game players. *Nat. Struct. Mol. Biol.* **18**, 1175–1177 (2011).
- Eiben, C. B. et al. Increased Diels–Alderase activity through backbone remodeling guided by Foldit players. *Nat. Biotechnol.* **30**, 190–192 (2012).
- Blout, E. R. & Idelson, M. Compositional effects on the configuration of water-soluble polypeptide copolymers of L-glutamic acid and L-lysine. *J. Am. Chem. Soc.* **80**, 4909–4913 (1958).
- Doty, P., Imahori, K. & Klemperer, E. The solution properties and configurations of a polyampholytic polypeptide: copoly-L-lysine-L-glutamic acid. *Proc. Natl Acad. Sci. USA* **44**, 424–431 (1958).
- Ghosh, K. & Dill, K. A. Theory for protein folding cooperativity: helix bundles. *J. Am. Chem. Soc.* **131**, 2306–2312 (2009).
- Rohl, C. A., Strauss, C. E. M., Misura, K. M. S. & Baker, D. Protein structure prediction using Rosetta. *Methods Enzymol.* **383**, 66–93 (2004).
- Koga, N. et al. Principles for designing ideal protein structures. *Nature* **491**, 222–227 (2012).
- Regan, L. & DeGrado, W. F. Characterization of a helical protein designed from first principles. *Science* **241**, 976–978 (1988).
- Harbury, P. B., Plecs, J. J., Tidor, B., Alber, T. & Kim, P. S. High-resolution protein design with backbone freedom. *Science* **282**, 1462–1467 (1998).
- Thomson, A. R. et al. Computational design of water-soluble α -helical barrels. *Science* **346**, 485–488 (2014).
- Jacobs, T. M. et al. Design of structurally distinct proteins using strategies inspired by evolution. *Science* **352**, 687–690 (2016).
- Ramachandran, G. N. & Sasisekharan, V. Conformation of polypeptides and proteins. *Adv. Protein Chem.* **23**, 283–438 (1968).
- Chen, V. B. et al. MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D* **66**, 12–21 (2010).
- Montelione, G. T. et al. Recommendations of the wwPDB NMR Validation Task Force. *Structure* **21**, 1563–1570 (2013).
- Zhang, Y. & Skolnick, J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.* **33**, 2302–2309 (2005).
- Santorio, M. M. & Bolen, D. W. Unfolding free energy changes determined by the linear extrapolation method. 1. Unfolding of phenylmethanesulfonyl α -chymotrypsin using different denaturants. *Biochemistry* **27**, 8063–8068 (1988).

Acknowledgements We thank all Foldit players for their gameplay contributions, and for feedback offered on the Foldit website (<https://fold.it>). We thank A. Kang, S. A. Rettie, C. Chow and L. Carter for help with experiments; D. Alonso, L. Goldschmidt, P. Vecchiato and D. Kim for computer support; and Rosetta@home (<https://boinc.bakerlab.org>) volunteers for computing resources. We thank G. Rocklin, V. Mulligan and other members of the Baker laboratory for discussions. This material is based on work supported by the National Science Foundation (NSF) Graduate Research Fellowship under grant no. DGE-1256082, NSF grant no. 1629879, National Institutes of Health (NIH) grant 1UH2CA203780, and NIH grants 1S10 OD018207 and 5R01 GM120574 (to G.T.M.) and HHMI (D.B.). The ALS-ENABLE beamlines are supported in part by the NIH, National Institute of General Medical Sciences, grant P30 GM124169-01. The Advanced Light Source is a DOE User Facility under contract no. DE-AC02-05CH11231. Foldit3 was a nominated target of the CASP COMMONS Community Outreach program.

Reviewer information Nature thanks Jérôme Waldispühl and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions B.K., Z.P., F.K., S.C. and D.B. designed the study. B.K., J.F., T.H., A.F., D.-A.S. and S.C. developed Foldit software tools. A. Boykov, R.D.E., S.K., T.N.-S. and L.W., along with the other Foldit players, designed all proteins. B.K., F.K., A.F. and A. Bauer analysed Foldit player designs. B.K. performed biophysical characterization. B.K., M.J.B. and F.D. determined crystal structures. G.L., Y.I. and G.T.M. determined the NMR structure. B.K. and D.B. wrote the manuscript with input from all authors. Foldit players contributed extensively through their feedback and gameplay, which generated the data for this paper.

Competing interests G.T.M. is a co-founder of Nexomics Biosciences.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-019-1274-4>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-019-1274-4>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to D.B.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

METHODS

Foldit protein design puzzles. Foldit puzzles were set up with a model polyisoleucine in fully extended conformation, with fixed length ranging from 60 to 100 residues. Each puzzle was posted online for seven days, during which Foldit players competed to develop a protein model with the lowest energy, as calculated by the Rosetta energy function. Foldit puzzles used the talaris2013_cart scorefunction with the following modifications: (1) the cart_bonded scoreterm was upweighted (increased from 0.5 to 2.0) to ensure realistic bond lengths and angles as players cut and splice the backbone chain; (2) a penalty-only envsmooth scoreterm (weighted at 2.0) was added to supplement the Rosetta solvation treatment, and to discourage the design of buried polar and exposed nonpolar residues; (3) the reference energy of alanine was modified (increased to 3.0) to discourage the excessive design of alanine. Configuration files for all Foldit puzzles are provided in the Supplementary Data. Each Foldit puzzle was accompanied by a brief description, along with an explanation of any supplementary rules enforced in the puzzle. Design puzzles were accessible to all Foldit users; Foldit user registration is free and open to the public, at <http://fold.it>. Models were collected continuously as Foldit players worked on the puzzles, as the Foldit application automatically uploads the user's latest model to a server every 2–5 min. This study was approved by the University of Washington Institutional Review Board, and informed consent for this research was obtained from all Foldit users at the time of user registration.

Protein design selection. After the end of each puzzle, we selected player models for further analysis as follows: first, we selected the lowest-energy model from each of the ten top-ranked groups, in which independent players were treated as individual groups (designs named with suffix '0000-9'). Second, we selected the lowest-energy model from the ten top-ranked solo players, which includes independent players as well as group members that developed a model without assistance from their group (suffix 's000-9'). Third, we visually inspected models that were flagged by Foldit players for special consideration, and selected any models that appeared plausible (suffix 'S***'). Last, we ranked and pruned the set of remaining models by removing any models that align to a better-scoring model with C_α r.m.s.d. less than 2.5 Å. We visually inspected the 50 top-ranked models in the pruned set and selected any models that appeared plausible (suffix '1001-50'). Models deemed 'implausible' typically lacked secondary structure, contained buried polar residues or included long stretches of completely polar residues. At each step, we used TM-align²⁶ to eliminate duplicate models (TM-score >0.98) that had already been selected (for example, models that were top-ranking and flagged by players). In rounds 2 and 3, the top-ranked group and solo models were automatically selected for further analysis, without visual inspection. The sequences of selected models were subjected to Rosetta ab initio structure prediction¹⁷, using the distributed computing platform Rosetta@home. If ab initio predictions identified any decoy structures with energy comparable to (or lower than) the designed structure, or if ab initio predictions were unable to sample the designed structure, the design was rejected. All other designs were selected for experimental characterization. See Extended Data Table 1 for summary statistics on design selection. The majority of experimentally tested designs (96 of 146) were top-ranked group or solo designs, which were selected 'blindly' (without visual inspection). Models and FASTA sequences of all tested designs are shown in the Supplementary Data.

Protein expression and purification. A 6×His tag with TEV-cleavable linker (sequence MGHHHHHHGWSENLYFQGS) was prepended to the N terminus of each design selected for experimental characterization. Plasmids containing the encoded genes were ordered from Genscript in pET15 (designs with prefix between 997258 and 1998925), in pET21 (1998555–2002990) or from Twist in pET29 (2003048–2003594) vectors. Plasmids were transformed into *E. coli* BL21 Star (DE3) cells (Invitrogen), and grown overnight in 4 ml Luria–Bertani (LB) medium with 50 µg/ml carbenicillin (for pET15, pET21 vectors) or 30 µg/ml kanamycin (for pET29). Overnight cultures were used to inoculate 0.5 l auto-induction medium, and grown at 37 °C for 18 h. Cultures were pelleted and resuspended in 25 ml lysis buffer (20 mM Tris pH 8.0, 300 mM NaCl, 1 mg/ml lysozyme, 0.1 mg/ml DNase, 1 mM PMSF), and lysed by microfluidization. The cell lysate was pelleted and supernatant was filtered with a 0.22-µm filter before loading onto a 2 ml nickel-affinity gravity column. Protein bound to the column was washed with 20 ml wash buffer (20 mM Tris pH 8.0, 500 mM NaCl, 30 mM imidazole) and eluted in 10 ml elution buffer (20 mM Tris pH 8.0, 500 mM NaCl, 250 mM imidazole). Purified protein was dialysed into TBS (20 mM Tris pH 8.0, 300 mM NaCl) at 4 °C overnight to remove imidazole and further purified by size-exclusion chromatography on an AKTApurify (GE Healthcare) with a Superdex S75 10/300 GL column (GE Healthcare). For proteins containing cysteine, dialysis and gel filtration were carried out in TBS with 1 mM tris(2-carboxyethyl)phosphine (TCEP). Protein expression and solubility was determined from SDS–PAGE and mass spectrometry. Oligomeric state was determined by size-exclusion chromatography.

Circular dichroism. Purified protein was dialysed into 50 mM sodium phosphate pH 7.4 at 4 °C overnight (plus 500 µM TCEP for proteins containing cysteine). All circular dichroism data were collected on an AVIV Model 420 spectrometer.

Far UV spectra and temperature melts were measured with 11–62 µM protein in a quartz cuvette with path length of 1 mm. Protein concentration was determined by absorbance at 280 nm using a NanoDrop spectrophotometer (Thermo Scientific), using predicted extinction coefficients. Wavelength spectra were measured between 195 and 260 nm at 25 °C, 95 °C and again after cooling to 25 °C. For temperature melts, ellipticity at 220 nm was monitored as temperature increased from 25 °C to 95 °C in increments of 2 °C. Chemical titrations were carried out with 1.0–21 µM protein in a quartz cuvette with path length of 10 mm. Ellipticity at 220 nm was monitored at concentrations of guanidinium chloride increasing from 0 to 7 M, in increments of 0.25 M. Denaturation curves were fitted with nonlinear regression to two-state unfolding model with six parameters: the folding free energy, m -value, and slope and y intercept for baseline curves²⁷.

X-ray crystallography. Prior to X-ray crystallography, the N-terminal 6×His tag was cleaved from protein samples by incubation with 250 µg TEV protease at 25 °C for 4 h in 20 mM Tris pH 8.0, 300 mM NaCl, 1 mM DTT. The reaction product was dialysed into TBS overnight at 4 °C to remove DTT and flowed over a 2 ml metal-affinity gravity column to remove TEV protease and residual histidine tag. The cleaved protein was further purified by gel filtration as described above. Purified protein was concentrated to 20–100 mg/ml in 20 mM Tris pH 8.0, 300 mM NaCl. Crystallization screening was carried out with a variety of 96-condition spare matrix suites available from Qiagen or Hampton Research. A Mosquito Crystal nanolitre robot (TTP Labtech) was used to prepare screens in 3-well sitting drop plates, with 200 nl drops and protein:precipitant ratios of 1:1, 1:2 and 2:1.

Foldit1 (2002949_0000) was crystallized at 20 mg/ml in 50 mM HEPES pH 7.5, 0.2 M potassium chloride, 35% v/v pentaerythritol propoxylate. Crystals were flash-frozen in liquid nitrogen without further cryo-protection. X-ray diffraction was collected to a resolution of 1.18 Å.

Peak6 (2003333_0006) was crystallized at 40 mg/ml in 0.1 M sodium acetate pH 4.5, 0.2 M lithium sulphate, 50% w/v PEG 400. Crystals were briefly soaked in mother liquor plus 20% PEG 200, then flash-frozen in liquid nitrogen. X-ray diffraction was collected to a resolution of 1.54 Å.

Ferredog–Diesel (2003169_S953) was crystallized with 6×His tag intact, at 80 mg/ml in 0.1 M citrate pH 4.0, 3.0 M NaCl. Crystals were dehydrated by soaking in 5 µl mother liquor in open air for 10 min, then flash-frozen in liquid nitrogen. X-ray diffraction was collected to a resolution of 1.92 Å.

X-ray diffraction datasets were collected at the Advanced Light Source (Berkeley, CA). Data was processed with HKL2000²⁸. Crystal structures were solved by molecular replacement with Phaser²⁹, using the backbone of the original designed model with sidechains truncated to the β-carbon (Foldit1 and Peak6), or using the backbone of a model predicted ab initio from the design sequence (Ferredog–Diesel). Models were built and refined in iterative cycles using Coot and PHENIX^{30,31}. Diffraction data and refinement statistics are listed in Extended Data Table 2.

NMR spectroscopy. NMR studies were performed using uniformly ¹⁵N, ¹³C-enriched protein samples. A synthetic gene for Foldit3 (2003265_s008) was obtained from Genscript already incorporated into plasmid pET15TEV_NESG, which includes a N-terminal 6×His purification tag, followed by a TEV protease cleavage site (sequence MGHHHHHHGWSENLYFQGS). *E. coli* BL21(DE3) cells containing plasmid pET15TEV_NESG-Foldit3 were grown in 1 l MJ9 minimal medium³², supplemented with 100 µg/ml ampicillin at 37 °C. To produce uniformly ¹⁵N and ¹³C-enriched protein samples, 1 g/l ¹⁵NH₄ salts and 2g/l U-¹³C glucose were added as sole nitrogen and carbon sources, respectively. When OD₆₀₀ reached around 0.5 units, the culture was transferred to 18 °C, and protein production was induced by addition of 1 mM IPTG. After overnight incubation, the cells were collected and resuspended in 20 ml binding buffer (20 mM Tris–HCl pH 8.0, 500 mM NaCl and 20 mM imidazole). After passing the cells through 900–1000 psi French press twice, cell debris were removed by 10,000 r.p.m. for 30 min. The supernatant was further spun down at 40,000 r.p.m. for 1 h. The obtained supernatant (soluble fraction) was mixed with 1 ml of Ni-resin and incubated at 4 °C for 1 h. The non-specific binding proteins were removed by 20 ml binding buffer and washing buffer (20 mM Tris–HCl pH 8.0, 500 mM NaCl and 50 mM imidazole) and the target protein was eluted by 5 ml elution buffer (20 mM Tris–HCl pH 8.0, 500 mM NaCl and 300 mM imidazole). The protein was dialysed against GF buffer (20 mM Tris–HCl pH 8.0, 100 mM NaCl) for overnight and gel filtration was carried out using AKTA express with high-load 26/600 Superdex 200 pg column. Homogeneity (>97%) was validated by SDS polyacrylamide gel electrophoresis. The purified protein was dialysed against 20 mM potassium phosphate (pH 6.5), and the protein concentration was adjusted to between 0.3–0.4 mM for NMR studies.

All NMR spectra were recorded at 25 °C using cryogenic NMR probes. All NMR data were collected on the Bruker AVANCE III 600 MHz spectrometers and processed using the program NMRPipe³³, and analysed using the programs SPARKY and XEASY³⁴. Spectra were referenced to external DSS. Sequence-specific resonance assignments were determined using AutoAssign software together with interactive manual analysis, as previously described³⁵. Backbone dihedral angle

constraints were derived from the chemical shifts using the program TALOS_N³⁶ for residues located in well-defined secondary structure elements. The programs ASDP³⁷ and CYANA^{38,39} were used to automatically assign NOEs and to calculate structures. RPF analysis^{37,40} was used in parallel to guide iterative cycles of noise and artefact peak removal, peak picking, and NOESY peak assignments. The 20 conformers with the lowest target CYANA function value were then refined in explicit water⁴¹ using the program CNS⁴². The structural statistics and global structure quality factors (Extended Data Table 3) including Verify3D⁴³, ProsaII⁴⁴, PROCHECK⁴⁵, and MolProbity⁴⁶ raw and statistical Z-scores were computed using the PSVS⁴⁷ v.1.5 and PDBStat⁴⁸ software packages. The global goodness-of-fit of the final structure ensembles with the NOESY peak list data, the NMR DP score, was determined using the RPF analysis program⁴⁰.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

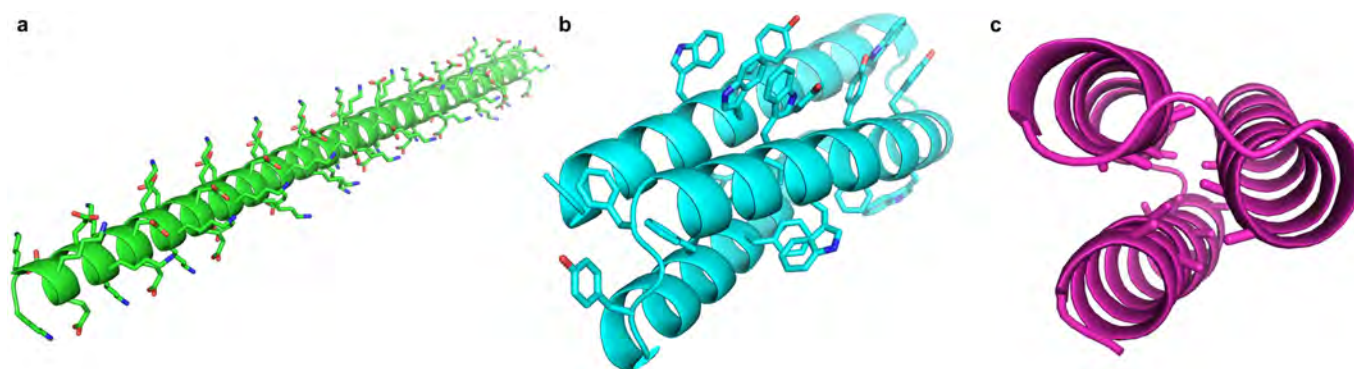
Data availability

The atomic coordinates of Foldit1, Peak6 and Ferredoxin-Diesel crystal structures and the Foldit3 NMR structure have been deposited in the RCSB Protein Data Bank (PDB) with accession numbers 6MRR, 6MRS, 6NUK and 6MSP, respectively. Chemical shift and NOESY peak list data for Foldit3 were deposited in the Biological Magnetic Resonance Data Bank with accession number 30527.

Code availability

Because Foldit crowdsourcing relies on regulated, fair competition between participants, the source code of the Foldit user interface is not open. The underlying Rosetta macromolecular modelling suite (<https://www.rosettacommons.org>) is freely available to academic and non-commercial users, and commercial licenses are available via the University of Washington CoMotion Express License Program. Analysis scripts used in this paper are available in the Supplementary Information.

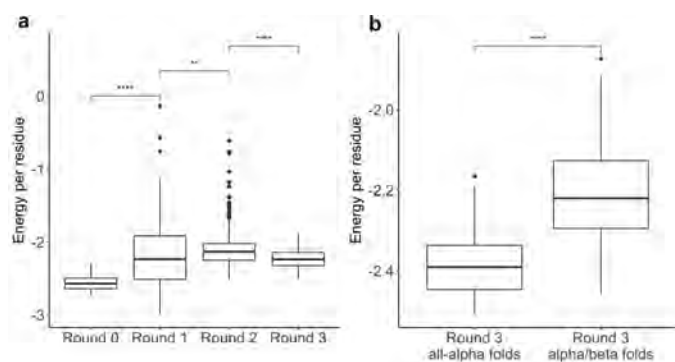
28. Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276**, 307–326 (1997).
29. McCoy, A. J. et al. Phaser crystallographic software. *J. Appl. Crystallogr.* **40**, 658–674 (2007).
30. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D* **66**, 486–501 (2010).
31. Afonine, P. V. et al. Towards automated crystallographic structure refinement with phenix.refine. *Acta Crystallogr. D* **68**, 352–367 (2012).
32. Jansson, M. et al. High-level production of uniformly ¹⁵N- and ¹³C-enriched fusion proteins in *Escherichia coli*. *J. Biomol. NMR* **7**, 131–141 (1996).
33. Delaglio, F. et al. NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J. Biomol. NMR* **6**, 277–293 (1995).
34. Bartels, C., Xia, T. H., Billeter, M., Güntert, P. & Wüthrich, K. The program XEASY for computer-supported NMR spectral analysis of biological macromolecules. *J. Biomol. NMR* **6**, 1–10 (1995).
35. Liu, G. et al. NMR data collection and analysis protocol for high-throughput protein structure determination. *Proc. Natl Acad. Sci. USA* **102**, 10487–10492 (2005).
36. Shen, Y., Delaglio, F., Cornilescu, G. & Bax, A. TALOS+: a hybrid method for predicting protein backbone torsion angles from NMR chemical shifts. *J. Biomol. NMR* **44**, 213–223 (2009).
37. Huang, Y. J., Tejero, R., Powers, R. & Montelione, G. T. A topology-constrained distance network algorithm for protein structure determination from NOESY data. *Proteins* **62**, 587–603 (2006).
38. Güntert, P., Mumenthaler, C. & Wüthrich, K. Torsion angle dynamics for NMR structure calculation with the new program DYANA. *J. Mol. Biol.* **273**, 283–298 (1997).
39. Herrmann, T., Güntert, P. & Wüthrich, K. Protein NMR structure determination with automated NOE assignment using the new software CANDID and the torsion angle dynamics algorithm DYANA. *J. Mol. Biol.* **319**, 209–227 (2002).
40. Huang, Y. J., Powers, R. & Montelione, G. T. Protein NMR recall, precision, and F-measure scores (RPF scores): structure quality assessment measures based on information retrieval statistics. *J. Am. Chem. Soc.* **127**, 1665–1674 (2005).
41. Linge, J. P., Williams, M. A., Spronk, C. A., Bonvin, A. M. & Nilges, M. Refinement of protein structures in explicit solvent. *Proteins* **50**, 496–506 (2003).
42. Brünger, A. T. et al. Crystallography & NMR system: a new software suite for macromolecular structure determination. *Acta Crystallogr. D* **54**, 905–921 (1998).
43. Lüthy, R., Bowie, J. U. & Eisenberg, D. Assessment of protein models with three-dimensional profiles. *Nature* **356**, 83–85 (1992).
44. Sippl, M. J. Recognition of errors in three-dimensional structures of proteins. *Proteins* **17**, 355–362 (1993).
45. Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. Procheck—a program to check the stereochemical quality of protein structures. *J. Appl. Crystallogr.* **26**, 283–291 (1993).
46. Word, J. M., Bateman, R. C., Jr, Presley, B. K., Lovell, S. C. & Richardson, D. C. Exploring steric constraints on protein mutations using MAGE/PROBE. *Protein Sci.* **9**, 2251–2259 (2000).
47. Bhattacharya, A., Tejero, R. & Montelione, G. T. Evaluating protein structures determined by structural genomics consortia. *Proteins* **66**, 778–795 (2007).
48. Tejero, R., Snyder, D., Mao, B., Aramini, J. M. & Montelione, G. T. PDBStat: a universal restraint converter and restraint analysis software package for protein NMR. *J. Biomol. NMR* **56**, 337–351 (2013).
49. Trifonov, E. N. in *Structure and Methods, Vol. 1: The Proceedings of the Sixth Conversation held at The University-SUNY (Adenine)*, 1990).
50. Holm, L. & Laakso, L. M. Dali server update. *Nucleic Acids Res.* **44** (W1), W351–W355 (2016).



Extended Data Fig. 1 | Initial top-ranking Foldit player designs.

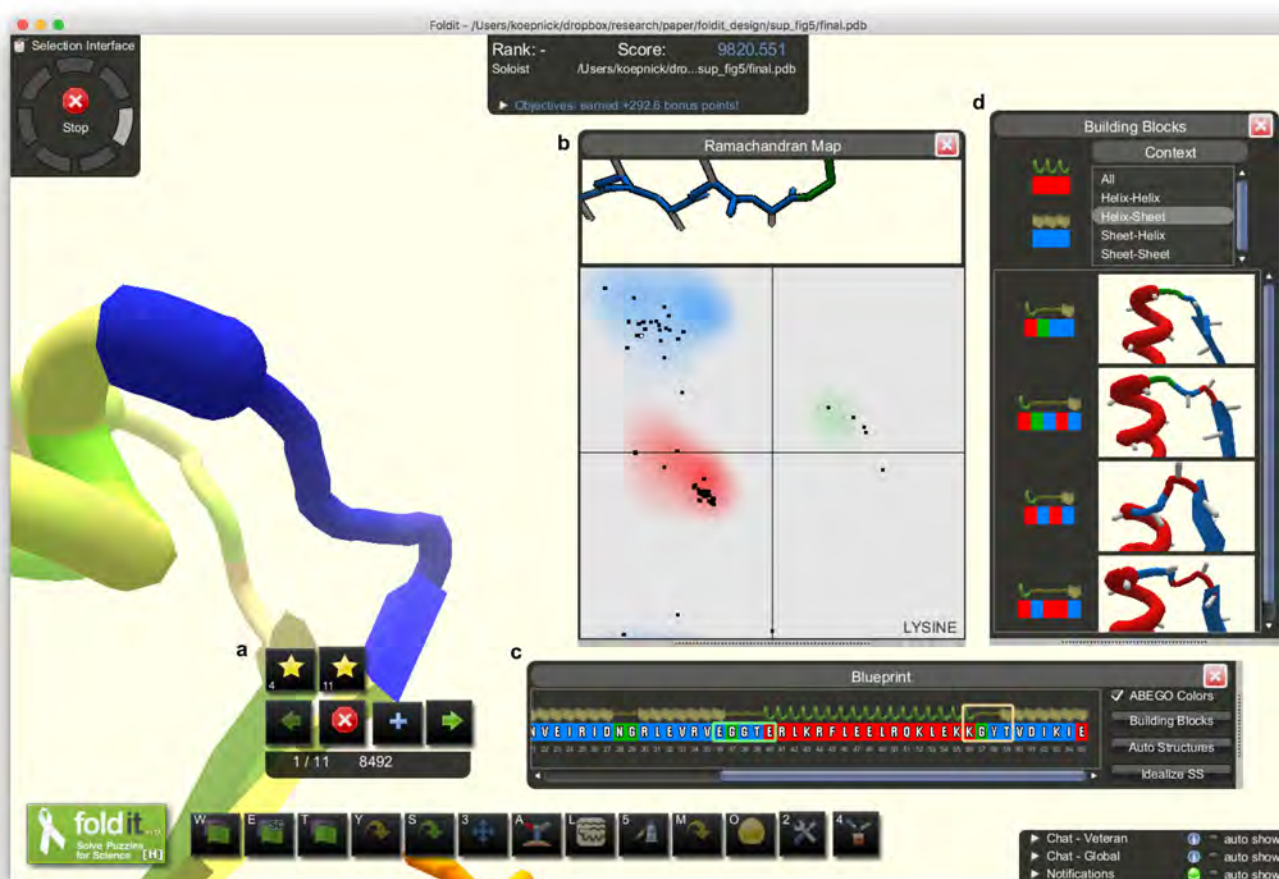
When challenged to design a protein with only the talaris2013 score function (and no additional rules), Foldit players discovered low-energy models that are unlikely to fold as designed. **a**, An extended α -helix, composed entirely of lysine and glutamate, has very favourable energies for hydrogen-bonding, electrostatic and backbone torsions, but is unlikely to fold cooperatively into a single stable structure. This type of design is discouraged with the 'core exists' rule. **b**, Owing to their greater surface area, large aromatic sidechains can make more interactions than

smaller aliphatic sidechains, even when underpacked or solvent-exposed. This type of design is discouraged with the 'residue interaction energy' rule. **c**, A design with an alanine- and glycine-saturated core can make favourable van der Waals interactions between closely packed backbone atoms; however, the burial of these small sidechains is associated with a weaker hydrophobic effect, and the lack of interdigitation allows exchange between multiple conformations with similar core packing energies (that is, molten globule behaviour). These designs are discouraged with the 'secondary structure design' rule.



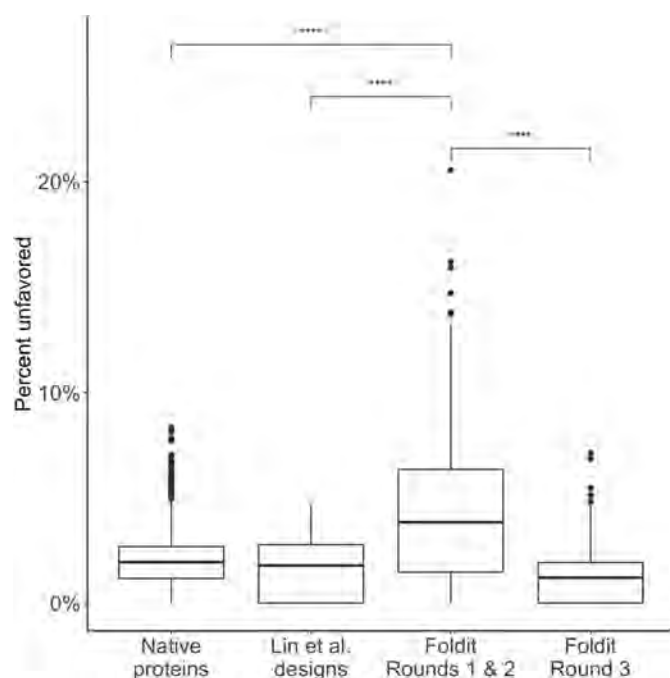
Extended Data Fig. 2 | Rosetta energy of top Foldit player designs.

Rosetta energy of top-ranking designs was calculated with the *talaris2013* score function and normalized by residue count. **a**, Energy of top-ten-ranked designs from: initial Foldit puzzles (round 0; $n = 30$ designs), round 1 puzzles ($n = 170$), round 2 puzzles ($n = 510$) and round 3 puzzles ($n = 250$). The introduction of supplementary rules in round 1 and round 2 resulted in higher-energy designs ($P < 10^{-6}$ and $P < 0.01$, respectively; Wilcoxon rank-sum test). The backbone modelling improvements in round 3 resulted in lower-energy designs ($P < 10^{-15}$; Wilcoxon rank-sum test). **b**, Energy of top-ten-ranked designs from round three all- α puzzles ($n = 30$) or α/β -puzzles using the 'secondary structure' rule ($n = 220$). All- α designs tend to have lower energy than α/β -designs ($P < 10^{-10}$; Wilcoxon rank-sum test). Box plots show: centre line, median; box limits, upper and lower quartiles; whiskers, $1.5 \times$ interquartile range; points, outliers.

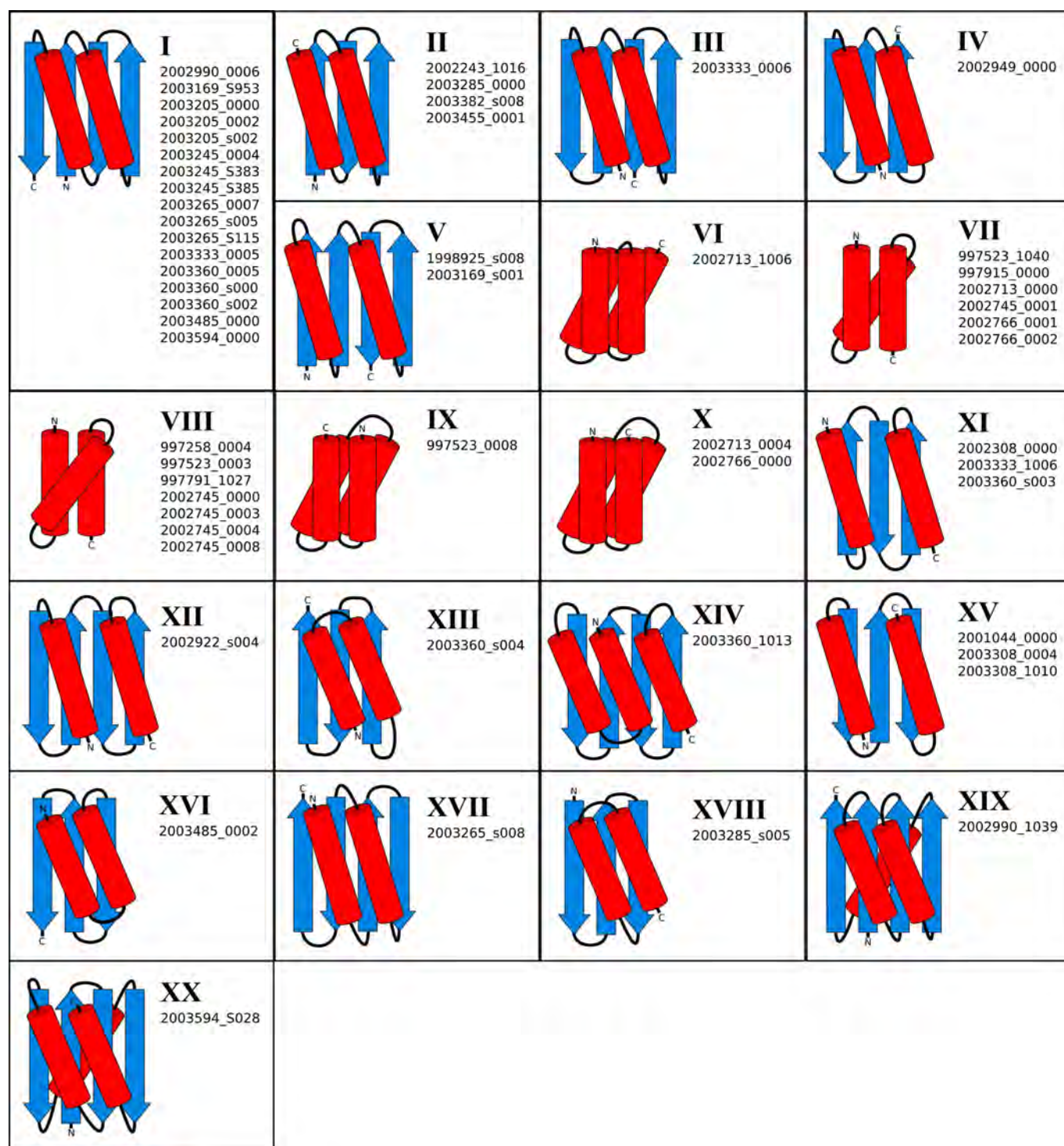


Extended Data Fig. 3 | New backbone-modelling tools in Foldit. **a**, The 'remix' tool allows players to select a region of the model and search a library of backbone fragments for a conformation that can be substituted. **b**, An interactive Ramachandran map allows players to easily identify residues with outlier backbone conformations. Players can also click and drag points on the Ramachandran map to set the backbone torsions of

individual residues. **c**, A 'blueprint' panel shows the primary sequence and secondary structure content of the model. Residues are coloured according to the ABEGO quadrants of the Ramachandran plot⁷. **d**, Players can drag-and-drop modular building blocks onto the blueprint panel to insert common turn conformations into their model.



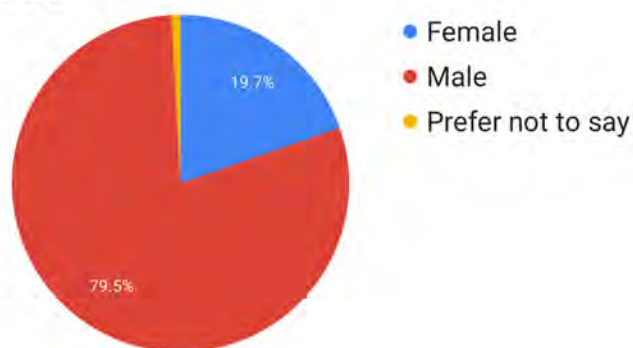
Extended Data Fig. 4 | Improvement of backbone quality in round 3 Foldit designs. MolProbity²⁴ was used to calculate the proportion of residues with unfavored or outlier backbone torsions in: high-resolution crystal structures of native proteins ($n = 6,342$), de novo design models from a previous study⁷ ($n = 72$), and top-ranking Foldit player designs from before ($n = 680$) and after ($n = 250$) improvements to Foldit backbone-modelling tools. Initial Foldit player designs contained significantly more unfavored torsions than native proteins or other de novo designs from a previous study⁷ ($P < 10^{-15}$, two-tailed t -test). Improvements to Foldit's backbone-modelling tools led Foldit players to produce designs with fewer unfavored torsions ($P < 10^{-15}$, two-tailed t -test). Box plots show: centre line, median; box limits, upper and lower quartiles; whiskers, $1.5 \times$ interquartile range; points, outliers.



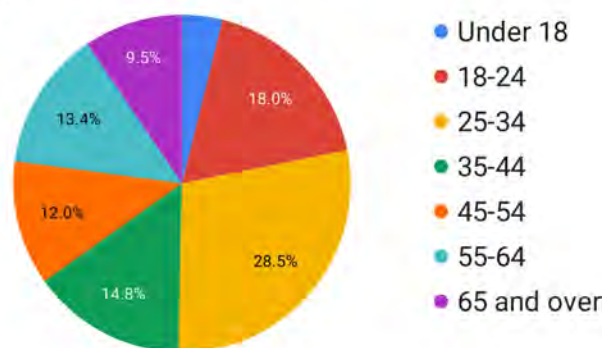
Extended Data Fig. 5 | Protein folds represented by successful Foldit player designs. Each fold has a unique arrangement and connectivity of secondary structure elements, depicted in cartoon diagrams. Diagrams are labelled with Roman numerals as in Fig. 3. Fold XX is a new fold,

previously unobserved in natural proteins; TM-align²⁶ and DALI⁵⁰ alignments of design 2003594_S028 against the entire PDB found no structural homologues with this fold.

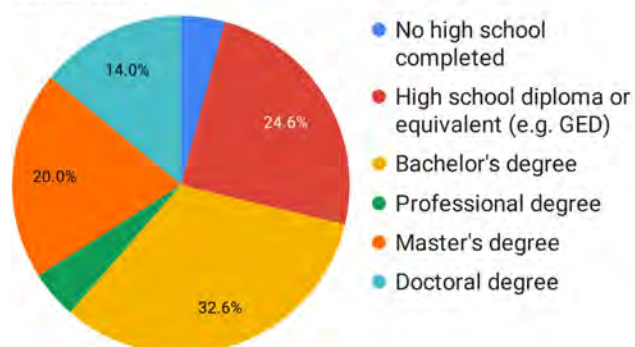
Gender



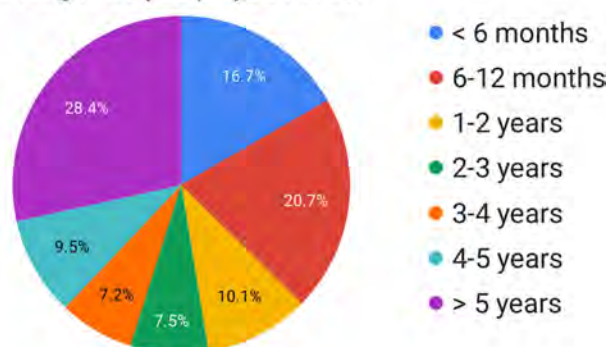
Age



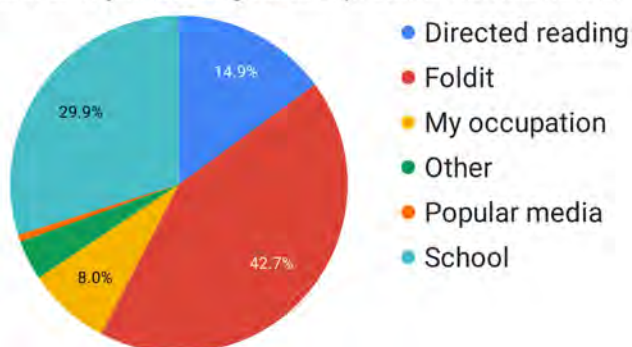
Education



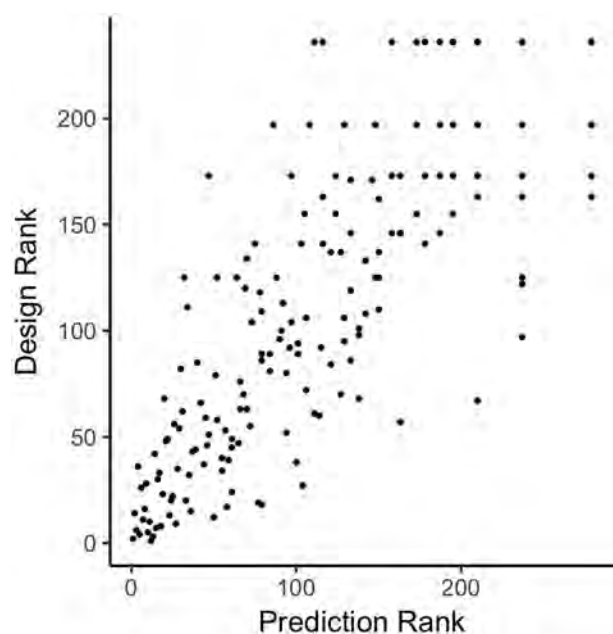
How long have you played Foldit?



Most of my knowledge about proteins comes from...



Extended Data Fig. 6 | Foldit player demographics. All players who participated in Foldit protein design puzzles and who had not opted out of Foldit-related email were solicited for survey questions. Data are shown for $n = 324$ responding Foldit players.



Extended Data Fig. 7 | Category rankings of Foldit players. Foldit player rankings are strongly correlated in the design and prediction categories (Spearman's rank correlation coefficient of 0.84). This suggests that skills developed playing Foldit structure prediction puzzles carry over to design puzzles and vice versa.

Extended Data Table 1 | Success rates of Foldit player-designed proteins

	Foldit player designs								Lin et al. ⁷
	Round 0		Round 1		Round 2		Round 3		
Sequence complexity*	0.20		0.35		0.44		0.21		0.20
Rosetta energy† (per residue)	-2.6	± 0.1	-2.2	± 0.5	-2.1	± 0.2	-2.2	± 0.1	-1.9 ± 0.1
Total puzzles	3		17		51		25		
Avg. players per puzzle	123	± 19	212	± 34	189	± 36	151	± 16	
Raw model count	140,273		2,887,213		10,556,093		4,124,471		
Top models	60		340		1020		500		
Shared models	53		214		726		342		
Clustered models	150		850		2550		1250		
Total models considered‡	263		1404		4296		2092		
Models selected for ab initio	0		100		1141		612		(Not reported)
Ab initio convergence	NA		12	12%	37	3%	99	16%	72
Models tested	NA		12		37		97		72
Expressed	NA		12	100%	23	62%	86	89%	70 97%
... and soluble	NA		12	100%	18	49%	71	73%	64 89%
... and monomeric	NA		7	58%	7	19%	52	54%	39 54%
... and structured	NA		6	50%	4	11%	46	47%	29 40%
Number of unique folds	NA		3		4		19		2

*Linguistic sequence complexity⁴⁹ was calculated from the top-ten-ranked models in all puzzles, using word lengths of 1, 2 and 3.

†Rosetta energy is the talaris2013 energy normalized by residue count. Values shown are mean and standard deviation for the ten top-ranked models in all puzzles. See Extended Data Fig. 2 for sample sizes.

‡Includes redundant models, as very similar models can appear in two or more categories (top, shared and clustered). See Methods for details on model selection.

Extended Data Table 2 | X-ray crystallography data and refinement statistics

	Foldit1 (6MRR)	Peak6 (6MRS)	Ferredog-Diesel (6NUK)
Data collection			
Space group	P 1 2 ₁ 1	P 3 ₁ 2 1	P 4 ₂ 2 ₁ 2
Cell dimensions			
<i>a</i> , <i>b</i> , <i>c</i> (Å)	24.05, 43.58, 29.28	52.41, 52.41, 56.09	69.21, 69.21, 90.59
α , β , γ (°)	90, 99.0, 90	90, 90, 120	90, 90, 90
Resolution (Å)	28.92 - 1.18 (1.222 - 1.18)	26.21 - 1.541 (1.596 - 1.541)	45.29 - 1.916 (1.985 - 1.916)
<i>R</i> _{merge}	0.02508 (0.1209)	0.0872 (0.7896)	0.08947 (3.164)
<i>I</i> / σ <i>I</i>	25.65 (9.97)	18.52 (1.34)	16.94 (0.86)
Completeness (%)	92.67 (88.38)	94.86 (65.00)	99.06 (97.65)
Redundancy	3.3 (3.4)	10.1 (4.8)	11.7 (11.5)
Refinement			
Resolution (Å)	1.18	1.541	1.916
No. reflections	18574	12861	17376
<i>R</i> _{work} / <i>R</i> _{free}	0.146 / 0.182	0.168 / 0.198	0.248 / 0.291
No. atoms			
Protein	574	646	1672
Ligand/ion	0	20	0
Water	116	89	37
<i>B</i> -factors			
Protein	14.54	22.82	69.09
Ligand/ion	0	47.36	0
Water	25.39	35.49	55.90
R.m.s. deviations			
Bond lengths (Å)	0.008	0.007	0.005
Bond angles (°)	0.83	1.03	1.01

Values in parentheses are for highest resolution shell. X-ray diffraction data for each protein structure were collected on a single crystal and processed as described in the Methods.

Extended Data Table 3 | NMR and refinement statistics for protein structures

	Foldit3 (6MSP)
NMR distance and dihedral constraints	
Distance constraints	
Total NOE	2012
Intra-residue	553
Inter-residue	
Sequential ($ i - j = 1$)	505
Medium-range ($ i - j < 4$)	301
Long-range ($ i - j > 5$)	653
Hydrogen bonds	66
Total dihedral angle restraints	
ϕ	59
ψ	59
Structure statistics	
Violations	
Distance constraints (Å)	0.01
Dihedral angle constraints (°)	0.88
Max. dihedral angle violation (°)	7.80
Max. distance constraint violation (Å)	0.66
Structure quality factors (raw score / Z-scores)	
Procheck G-factor (phi/psi only)	-0.09 / -0.04
Procheck G-factor (all dihedrals angles)	-0.14 / -0.83
Verify3D	0.45 / -0.16
Prosall (-ve)	0.91 / 1.08
MolProbity clashscore	17.51 / -1.48
Average pairwise r.m.s. deviation* (Å)	
Heavy	1.52
Backbone	0.71

*Pairwise r.m.s.d. was calculated among 20 refined structures for 'well-defined' residues 21–45, 48–54, 58–76, 81–87 and 90–96.

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a Confirmed

- ☒ ☒ The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- ☒ ☐ An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- ☐ ☒ The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- ☒ ☐ A description of all covariates tested
- ☒ ☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- ☒ ☐ A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- ☐ ☒ For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
Give P values as exact values whenever suitable.
- ☒ ☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- ☒ ☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- ☒ ☐ Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated
- ☒ ☐ Clearly defined error bars
State explicitly what error bars represent (e.g. SD, SE, CI)

Our web collection on [statistics for biologists](#) may be useful.

Software and code

Policy information about [availability of computer code](#)

Data collection

The pre-compiled Foldit game is freely available for download at <https://fold.it> for Windows, Linux, and Mac. A standalone version of Foldit is also freely available for academic use; for details visit <https://fold.it/standalone>. Foldit configuration files for all design puzzles are included in the Supplementary Information. The Rosetta software suite was used to perform ab initio prediction calculations; Rosetta is freely available for academic users on Github, and can be licensed for commercial use by the University of Washington CoMotion Express License Program.

Data analysis

Custom Python scripts written to analyze circular dichroism data are included in the Supplementary Information. Protein structures were analyzed with MolProbity (version 4.2). Crystallographic data were analyzed with PHENIX (release 1.101.1-2155) and Coot (v0.8.7 EL). NMR data were analyzed with SPARKY, XEASY, TALOS_N, ASDP, CYANA, PDBStat and PSVS (version 1.5).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The atomic coordinates of Foldit1, Peak6, and Ferredox-Diesel crystal structures, and the Foldit3 NMR structure, have been deposited in the RCSB Protein Data Bank with accession numbers 6MRR, 6MRS, 6NUK and 6MSP, respectively. Chemical shift and NOESY peak list data for Foldit3 were deposited in the Biological Magnetic Resonance Bank (BMRB ID 30527).

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample size for protein characterization was determined by estimated work load. In total, 146 protein designs were tested, from 97 Foldit puzzles. This was deemed sufficient due to the high number of successfully folded designs in our testing. For in silico designed-backbone analysis, the sample sizes of (n = 717 or 250) was considered sufficient. The inclusion of additional samples is not expected to affect the distribution of measured values.
Data exclusions	No data were excluded from analysis.
Replication	Puzzle configurations were used repeatedly in replicated Foldit puzzles to ensure reproducibility. The final puzzle configuration was used for 25 replicate Foldit puzzles. Protein expression and solubility was tested in duplicate. Structural characterization were performed once or twice with internal statistical validation. If positive results (e.g. protein expression, solubility, etc.) could not be replicated, they are reported as negative results.
Randomization	There was no randomized sample allocation in this work. All tested protein designs received identical treatment.
Blinding	Blinding was not relevant to this work, since all tested proteins received identical treatment.

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Unique biological materials
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input type="checkbox"/>	<input checked="" type="checkbox"/> Human research participants

Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

Human research participants

Policy information about [studies involving human research participants](#)

Population characteristics	Participation was open and free to the public, and we did not control for participant demographics. See Extended Data Fig. 6 for demographic data from a voluntary (non-obligatory) participant survey.
Recruitment	Participation was open and free to the public, at https://fold.it .

Heterochromatin drives compartmentalization of inverted and conventional nuclei

Martin Falk^{1,8}, Yana Feodorova^{2,3,8}, Natalia Naumova^{4,5,8}, Maxim Imakaev¹, Bryan R. Lajoie^{4,6}, Heinrich Leonhardt³, Boris Joffe³, Job Dekker⁴, Geoffrey Fudenberg^{1,7*}, Irina Solovoi^{3*} & Leonid A. Mirny^{1*}

The nucleus of mammalian cells displays a distinct spatial segregation of active euchromatic and inactive heterochromatic regions of the genome^{1,2}. In conventional nuclei, microscopy shows that euchromatin is localized in the nuclear interior and heterochromatin at the nuclear periphery^{1,2}. Genome-wide chromosome conformation capture (Hi-C) analyses show this segregation as a plaid pattern of contact enrichment within euchromatin and heterochromatin compartments³, and depletion between them. Many mechanisms for the formation of compartments have been proposed, such as attraction of heterochromatin to the nuclear lamina^{2,4}, preferential attraction of similar chromatin to each other^{1,4–12}, higher levels of chromatin mobility in active chromatin^{13–15} and transcription-related clustering of euchromatin^{16,17}. However, these hypotheses have remained inconclusive, owing to the difficulty of disentangling intra-chromatin and chromatin–lamina interactions in conventional nuclei¹⁸. The marked reorganization of interphase chromosomes in the inverted nuclei of rods in nocturnal mammals^{19,20} provides an opportunity to elucidate the mechanisms that underlie spatial compartmentalization. Here we combine Hi-C analysis of inverted rod nuclei with microscopy and polymer simulations. We find that attractions between heterochromatic regions are crucial for establishing both compartmentalization and the concentric shells of pericentromeric heterochromatin, facultative heterochromatin and euchromatin in the inverted nucleus. When interactions between heterochromatin and the lamina are added, the same model recreates the conventional nuclear organization. In addition, our models allow us to rule out mechanisms of compartmentalization that involve strong euchromatin interactions. Together, our experiments and modelling suggest that attractions between heterochromatic regions are essential for the phase separation of the active and inactive genome in inverted and conventional nuclei, whereas interactions of the chromatin with the lamina are necessary to build the conventional architecture from these segregated phases.

To test mechanisms of genome compartmentalization, we performed Hi-C in four mouse cell types that were isolated from primary tissues. These cell types have either conventional or inverted nuclear architectures: rod photoreceptors (inverted), non-rod retinal neurons (conventional), wild-type thymocytes (conventional) and lamin B receptor-null (*Lbr*^{−/−}) thymocytes^{20,21} (inverted) (Fig. 1a); data were collected from two biological replicates for each tissue type (Extended Data Fig. 1 and Supplementary Table 1). The latter three cell types provide points of comparison to rods: retinal non-rod neurons are similarly post-mitotic cells but have large conventional nuclei; wild-type and *Lbr*^{−/−} thymocytes are actively cycling cells with nuclei of a size similar to the nuclei of rods. Nuclear inversion of *Lbr*^{−/−} thymocytes is incomplete, most likely owing to regular cell divisions (Extended Data Fig. 2). Despite the large differences in nuclear organization that are evident from

microscopy (Fig. 1a), all features of chromatin organization characteristic of conventional nuclei—topologically associating domains (TADs), chromosome territories and compartments—are present in inverted nuclei, although with quantitative differences (Fig. 1b and Extended Data Figs. 3, 4, 5), as has also been seen recently in single-cell Hi-C²².

We subsequently investigated whether major differences in spatial positioning of euchromatin and heterochromatin affect nuclear compartmentalization as seen in Hi-C. We computed compartment profiles from Hi-C maps²³ (Fig. 1b) and defined the degree of compartmentalization as the enrichment of contacts between compartments of the same type (Methods). Although assignments of individual regions to euchromatic (A) and heterochromatic (B) compartments generally depend on the cell type, compartment profiles are highly correlated before and after perturbing the association of chromatin with the lamina in thymocytes—approaching the correlation between biological replicates (Extended Data Fig. 5b–e). The degree of compartmentalization decreases only slightly in thymocytes upon inversion, but becomes stronger in rods (Fig. 1c and Extended Data Fig. 5f). Together, our analyses show that the degree of compartmentalization is preserved despite the altered spatial positioning of individual A or B compartments upon inversion (Fig. 1a, d), and suggest that mechanisms of compartmentalization cannot be strictly dependent on the nuclear lamina.

To reconcile the similar Hi-C compartmentalization of inverted and conventional nuclei with the different spatial geometries in these nuclei, we sought a mechanism of compartmentalization that satisfied the three following criteria. First, it should reproduce the inverted organization, defined quantitatively with microscopy by the radial positions of different types of chromatin and with Hi-C by the strength of compartmentalization. Second, it should reproduce the conventional organization when attractive interactions between heterochromatin and the nuclear lamina are introduced. The conventional organization is characterized by a similar degree of compartmentalization in Hi-C, but a markedly different spatial location of compartments in microscopy. Third, it should be based on forces that are biologically and physically plausible. This limited us to short-range attractions between different chromatin types and of chromatin to the nuclear lamina.

To test mechanisms of compartmentalization, we developed an equilibrium polymer model of chromatin that represents chromosomes as block copolymers (Fig. 2a), similar to other phase-separation models of compartmentalization^{4–7}. Extending previous two-type models, our simulations use three types of monomers: euchromatin (A), heterochromatin (B) and pericentromeric constitutive heterochromatin (C). We modelled eight chromosomes—each consisting of 6,000 monomers; each monomer representing 40 kb of chromatin—confined to a spherical nucleus at 35% volume density²⁴. The sequence of A and B monomers along the polymer mirrors the sequence of compartments derived from Hi-C data of rods (Fig. 2a and Methods). To represent the satellite repeats of a pericentromeric region²⁵—or chromocentre—which

¹Institute for Medical Engineering and Science, and Department of Physics, Massachusetts Institute of Technology, Cambridge, MA, USA. ²Department of Medical Biology, Medical University-Plovdiv, Plovdiv, Bulgaria. ³Biozentrum, Ludwig Maximilians University Munich, Planegg-Martinsried, Germany. ⁴Howard Hughes Medical Institute, and Program in Systems Biology, Department of Biochemistry and Molecular Pharmacology, University of Massachusetts Medical School, Worcester, MA, USA. ⁵Epinomics Inc, Menlo Park, CA, USA. ⁶Illumina Inc, San Diego, CA, USA. ⁷Gladstone Institutes of Data Science and Biotechnology, University of California, San Francisco, San Francisco, CA, USA. ⁸These authors contributed equally: Martin Falk, Yana Feodorova, Natalia Naumova.

*e-mail: geoff.fudenberg@gmail.com; irina.solovoi@lrz.uni-muenchen.de; leonid@mit.edu

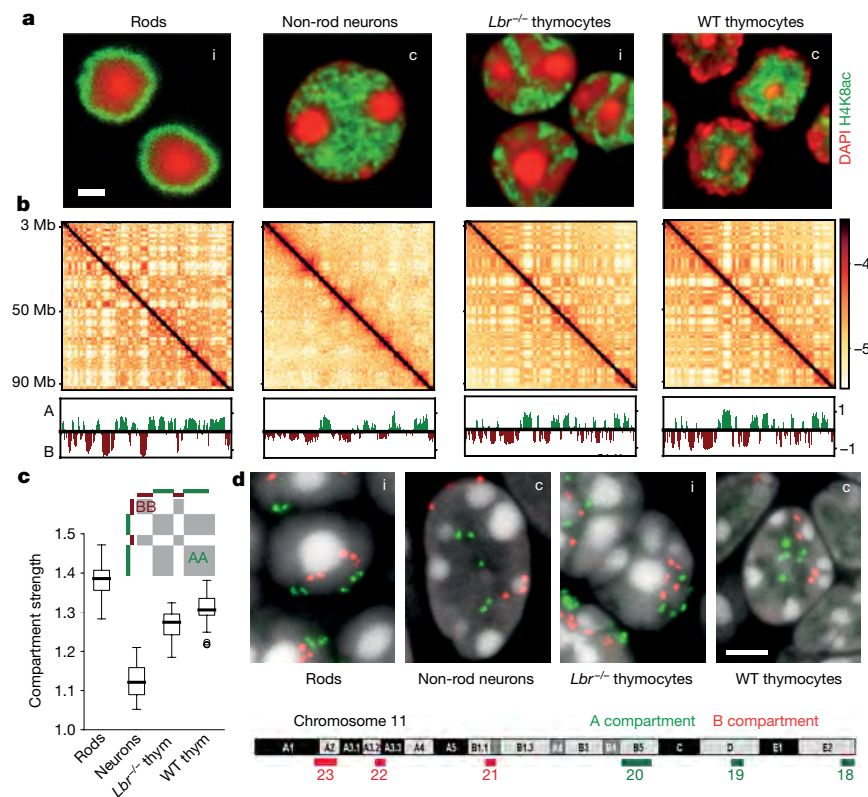


Fig. 1 | Microscopy and Hi-C analysis of conventional and inverted nuclei. **a**, Nuclei of non-rod neurons and wild-type (WT) thymocytes are conventional (c) with euchromatin residing in the interior. Rod nuclei are inverted (i) with a single central heterochromatic region (including the chromocentre) and euchromatin forming the peripheral shell. Nuclei of *Lbr*^{-/-} thymocytes are partially inverted and have several chromocentres. Euchromatin staining with anti-H4K8ac antibody (green); counterstain with DAPI (red), highlighting heterochromatin. Images are single optical sections. Scale bar, 2 μ m. See Extended Data Figs. 9a, 10a for schematic of positioning of euchromatin, heterochromatin and chromocentres. **b**, Hi-C contact maps (\log_{10} (contact frequency)) for an 87-Mb region of chromosome 1 (mm9) and corresponding compartment profiles indicating regions in the A (green) and B (red) compartment (see also Extended Data Fig. 1). Maps are corrected using ICE²³, with the matrix sums normalized to 1 (Methods). **c**, Compartmentalization is

strongest in rods and weakest in non-rod neurons; schematic indicates how compartmentalization is quantified ((AA + BB)/total). Box plots show compartmentalization calculated separately for each autosome in two replicates. Centre line is the median, the box ranges from the lower to upper quartiles and whiskers extend to $1.5 \times$ the interquartile range (see also Extended Data Fig. 5). **d**, Flipped localization of A and B compartment loci on chromosome 11 in inverted compared to conventional nuclei. Positions of detected compartments are marked with green (A compartment) and red (B compartment) bars below the chromosome ideogram. Fluorescence in situ hybridization (FISH) with a bacterial artificial chromosome (BAC) cocktail probe; BAC numbers are indicated below the compartment loci. Note the chromocentres are shown as bright globules in DAPI staining. Images are projections of 3- μ m confocal stacks. Scale bar, 2 μ m. The experiment was repeated twice.

is unmappable by Hi-C, we place a block of C monomers (16% of the chromosome length) at the proximal end of each chromosome. All monomers have excluded volume, and experience short-range pairwise attraction depending on their chromatin type. Given six pairwise attraction parameters (A–A, A–B, B–B, B–C, C–C and A–C), all possible permutations of attraction strengths specify 720 (6!) classes of models (see Methods). To constrain the space of possible models, we first quantitatively compared all 720 classes of models to microscopy data. Specifically, we computed the radial distributions for A, B and C monomers, and compared the distributions obtained in simulations with those obtained in microscopy¹⁹ (Fig. 2b and Methods).

Most model classes do not agree with the concentric geometry of the inverted nucleus observed in microscopy (Fig. 2c). For example, overly strong B–C interactions cause B and C to mix (Fig. 2c, model 8 and Extended Data Fig. 6a–c), while relatively weak B–C interactions lead to the expulsion of the C monomer chromocentres from a central mass of B monomers (Fig. 2c, model 112). Overly strong A–A interactions tend to encourage the formation of large euchromatic globules (Fig. 2c, model 650 and Extended Data Fig. 6d–f). Notably, this result argues against activity-related clustering of euchromatic regions^{13–16} as the main mechanism that underlies compartmentalization.

Only eight classes of models could reproduce the experimentally observed inverted geometry (Fig. 2b, c). These eight classes follow a

particular ordering of interaction strengths, which are, on average, dominated by heterochromatic interactions: A–A \approx A–B < A–C < B–B < B–C < C–C (Fig. 2d). We focused on the best-fitting class of models and further simplified these models by fixing C–C to be high enough to induce a central globule of C monomers, A–A to always be much smaller than B–B (Extended Data Fig. 7d), and all cross-terms to be the geometric means of the respective pure terms (for example, A–B = (A–A \times B–B)^{1/2}), thus satisfying the Flory–Huggins phase separation criterion²⁶. This leaves the B–B attraction as the only free parameter.

We next tested whether the heterochromatin-dominated models that reproduced the inverted organization seen in microscopy images could simultaneously reproduce the compartmentalization observed in Hi-C data. Fixing the order of interaction strengths, we found a range of the B–B attraction energies for which models could quantitatively reproduce both Hi-C and microscopy data (Fig. 3a, b). The central role of attractions between heterochromatic regions revealed by our analyses of inverted nuclei contrasts with suggestions that hinge on the importance of interactions between euchromatic regions^{13–16} or with the lamina² as the main drivers of compartmentalization. Stronger attractions between heterochromatic regions is consistent with the recently observed dominant role of heterochromatin-associated histone methylation in determining the mechanical properties of chromosomes²⁷.

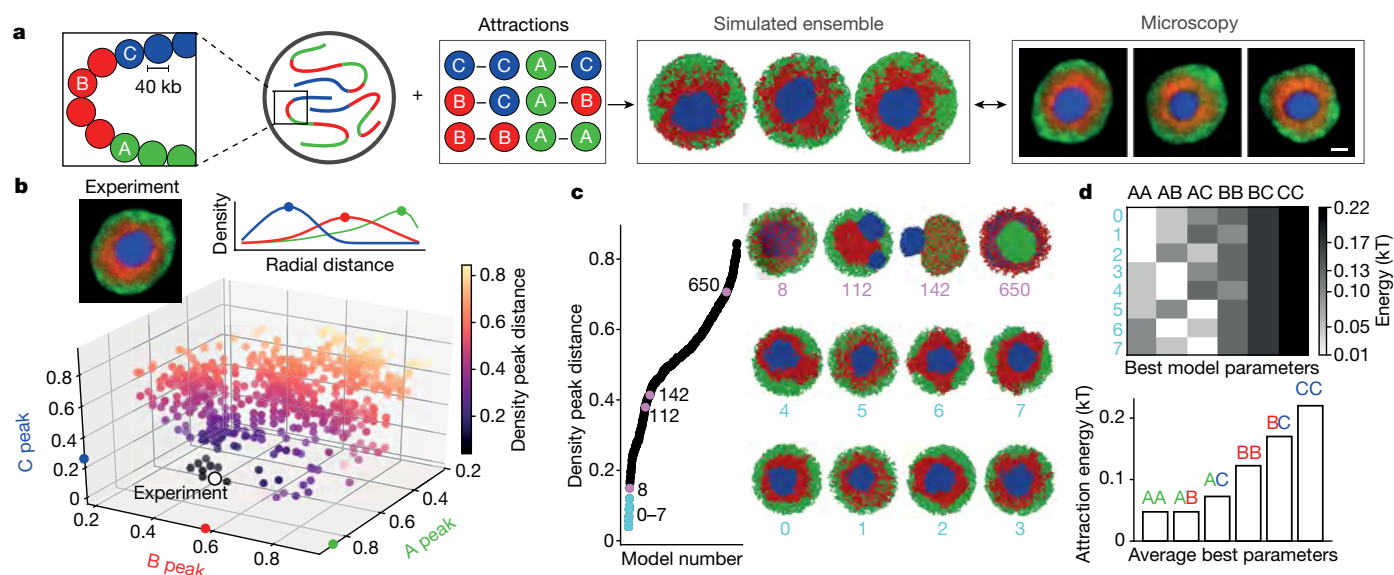


Fig. 2 | Morphology of the inverted nucleus restricts possible models of compartmentalization. **a**, Our approach was to first define mechanistic polymer models with parameters that describe the chromatin interactions between three types of monomers (A for euchromatin, B for heterochromatin and C for constitutive heterochromatin). Second, we simulated an ensemble of conformations for each model via Langevin dynamics. Finally, we compared simulations with experiments. To compare to microscopy, we computed radial distributions of A, B and C monomers. Models are characterized by relative attraction strengths between every pair of monomer types, leading to 720 (6!) classes of models. For analysis and other models, see Extended Data Fig. 6. Scale bar, 1 μm . **b**, Quantitative comparison of 720 model classes with microscopy

using the density peak distance, measuring the Euclidean distance between the peaks of the radial distributions for each chromatin class in simulations and experiment (white dot). Simulated densities computed from 50 configurations, experimental data from 24 nuclei¹⁹. **c**, Arranging the 720 models according to agreement with experimental data (that is, density peak distance; see Methods). The best eight models (0–7) are indicated in cyan. Other models are plotted in black, or pink, if a representative conformation is shown from that model. Models 8–15 are shown in Extended Data Fig. 6a. **d**, Heat map (top, individual models) and bar plot (bottom, averaged) of the best eight model parameters show that they increase on average as $A-A \approx A-B < A-C < B-B < B-C < C-C$.

To extend our model to conventional nuclei, we represented heterochromatin–lamina interactions with a short-ranged attraction^{2,20} (B–Lam attraction, Fig. 3c, d). To model the distinct chromocentres that were found experimentally, we pinned C-monomer clusters to random positions along the lamina. Pinning is not necessary to maintain distinct chromocentres for a period of time, but is needed to keep them separated in equilibrium simulations (Supplementary Video 1). By sweeping B–B and B–Lam attractions, we found that our model could simultaneously reproduce both the spatial positioning of active and inactive chromatin as observed in microscopy images, and the compartmentalization observed in Hi-C data for wild-type thymocytes. Whereas reproducing microscopy data requires sufficiently strong B–Lam without further constraining these parameters, simultaneously reproducing the compartmentalization observed in Hi-C data narrows down the range of B–Lam and B–B attractions (Fig. 3c, d). Notably, the region of best-fitting B–B attraction for conventional nuclei includes the best-fitting B–B attraction for inverted nuclei. As histone modifications remain associated with the same type of chromatin in inverted and conventional nuclei^{20,21}, we parsimoniously assume that B–B attraction remains the same in both nuclear types. With this constraint, we can narrow the range of possible B–Lam values (approximately 0.3 kT; Fig. 3c) and find that B–Lam attraction should be comparable to B–B attraction. Together, our simulations indicate that compartmentalization in both inverted and conventional nuclei is primarily controlled by heterochromatin–heterochromatin attractions, whereas heterochromatin–lamina attraction controls the global spatial morphology.

To test our proposed mechanism of compartmentalization, we simulated a time course of nuclear inversion (Fig. 4a, b). For this, we turned off lamina–heterochromatin interactions in simulated conventional nuclei and observed spontaneous inversion (Fig. 4b). Notably, the simulated time course mirrored key events during rod differentiation *in vivo*^{19,20} (Fig. 4b). B and C monomer droplets underwent irreversible liquid-like fusion in simulations, similar to other phase-separated systems^{10,11} (Fig. 4b, Extended Data Fig. 8a, c and Supplementary Video 2).

In simulations, although compartmentalization transiently dips after heterochromatin moves away from the lamina (Fig. 4a), compartments remain separated during the whole process of inversion. Consistently, microscopy shows that individual genomic loci reposition along with chromatin of their own compartment type during the entire process of nuclear inversion in rods *in vivo* (Extended Data Fig. 9). For example, the rhodopsin locus remains associated with euchromatin (A compartment) and the rhodopsin receptor remains expressed throughout the process of inversion (Fig. 4c).

To further test our proposed mechanism of compartmentalization, we initialized simulations from an inverted geometry and reintroduced lamina–heterochromatin interactions. These simulations predicted only partial de-inversion: whereas B monomers replaced A monomers at the periphery of the nucleus, C monomers remained as a single large globule surrounded by B monomers and associated with the lamina (Extended Data Fig. 10a). We tested these predictions experimentally by imaging de-differentiating rods of R7E mice²⁸ that express poly(Q)-expanded ataxin-7. Rods in these mice start lamin A/C expression after their nuclear inversion is completed²⁰ and acquire partially de-inverted morphologies that are remarkably similar to simulations (Extended Data Fig. 10b).

Together, our results show the central role of interactions between heterochromatin in establishing compartmentalization by phase separation. Using polymer simulations to reconcile microscopy and Hi-C data, we find that: (i) interactions between heterochromatic regions lead to phase separation of chromatin and these are essential for the compartmentalization of conventional and inverted nuclei; (ii) euchromatic interactions are dispensable for compartmentalization; and (iii) although lamina–heterochromatin interactions are dispensable for the segregation of euchromatin and heterochromatin, they are necessary to establish the conventional nuclear architecture. Although we narrow the search for key molecular determinants of compartmentalization to heterochromatin-associated molecules, making predictions for perturbations to particular molecular determinants

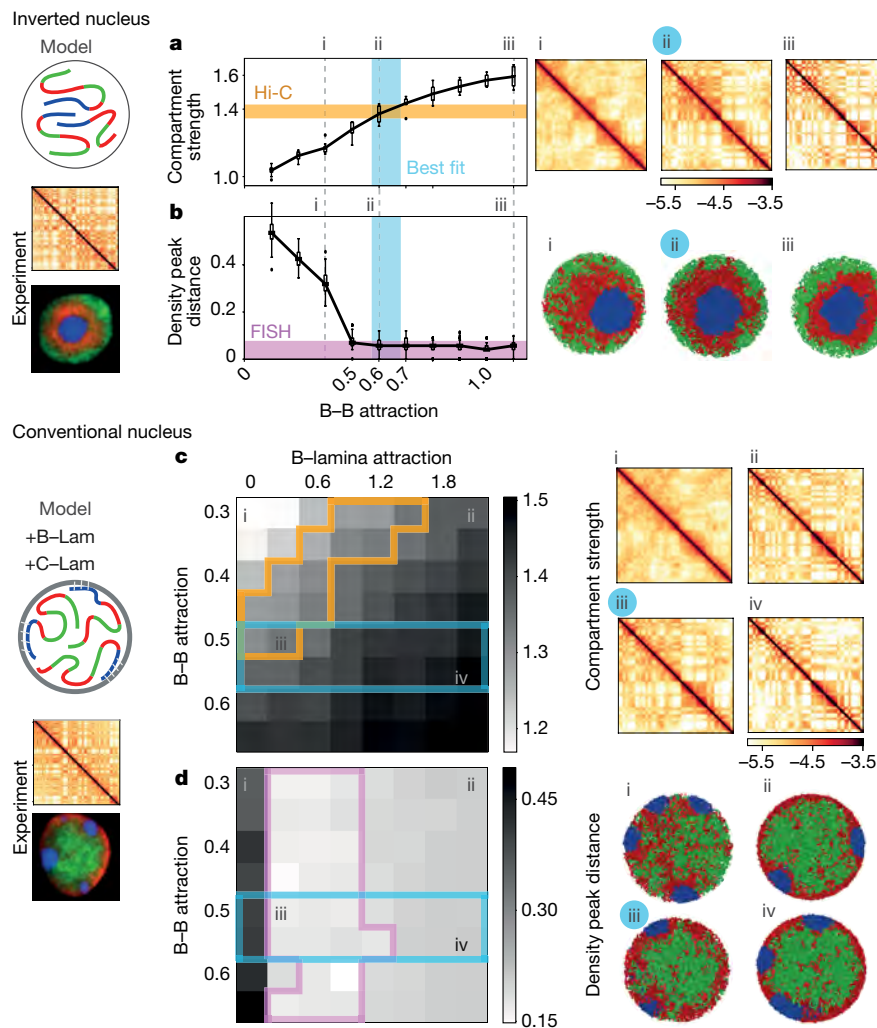


Fig. 3 | Heterochromatin-based mechanisms quantitatively reproduce inverted and conventional nuclei. **a, b,** Model for the inverted nucleus. Starting with the parameter ordering required to reproduce the morphology of the inverted nucleus (Fig. 2), we then varied B–B interactions to find models that best agree with Hi–C and microscopy data. **a,** Compartment strength as a function of B–B attraction (box plots as in Fig. 1c, with eight simulated chromosomes averaged across 150 conformations). Orange lane, compartment strength from rod Hi–C (see Fig. 1c). Blue, parameter range in agreement with Hi–C. Grey dashed lines, B–B values for which configurations and Hi–C maps are shown. i–iii, Simulated Hi–C maps ($\log_{10}(\text{contact frequency})$, chromosome 1: 50 Mb–150 Mb) are shown for indicated values of B–B. Model ii (highlighted) agrees best with Hi–C compartment strength. Attracting a small number of B monomers to the nuclear periphery does not disrupt the inverted architecture (Extended Data Fig. 7a). **b,** Distance between model and microscopy (as in Fig. 2b, c) as a function of B–B attraction (averaged over 150 conformations, box plots as in Fig. 1c). Purple lane, region of best agreement with microscopy (Methods); blue, as in **a**.

remains a limitation of our current study. Candidates for mediators of heterochromatin–heterochromatin interactions include affinity between homotypic repetitive elements^{1,9,29} or modified histones, and heterochromatin-associated proteins (for example, HP1)^{10,11}. Future work should consider the interplay between the mechanisms considered here and other chromosomal processes, such as non-equilibrium decondensation after mitosis³⁰ and loop extrusion⁸. Our results indicate that the inverted nucleus conceptually represents the default nuclear architecture imposed by the mechanism of compartmental interactions and that the conventional nucleus requires additional lamina–heterochromatin interactions. As most eukaryotic

Representative conformations are shown to the right (i–iii). **c, d,** Model for the conventional nucleus. The model for conventional nuclei additionally includes interactions of monomers with the nuclear lamina. B monomers are attracted to the lamina with a strength B–Lam and C monomer clusters are pinned to the lamina at random positions. **c,** Compartment strength as a function of B–B and B–Lam attractions, calculated as in **a**. i–iv, Simulated Hi–C maps displayed for indicated parameters. Experimental compartment strength (orange outline, for conventional wild-type thymocytes) can be matched (iii) even if B–B interactions are constrained to be the same as for inverted nuclei (blue outline). **d,** Distance between microscopy and models (over 150 simulated conformations). i–iv, Representative conformations for indicated parameters. Agreement with microscopy (purple outline) and Hi–C (blue outline) is simultaneously achievable (iii, highlighted) with B–B attraction strength from the best inverted nucleus model. Attracting a small number of A monomers to the periphery, or tethering a fraction of chromocentres to the interior, does not alter our conclusions (Extended Data Fig. 7).

nuclei have a conventional organization, our work raises questions about the functional relevance of heterochromatin positioning at the nuclear periphery.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-019-1275-3>.

Received: 11 January 2018; Accepted: 26 April 2019;
Published online 5 June 2019.

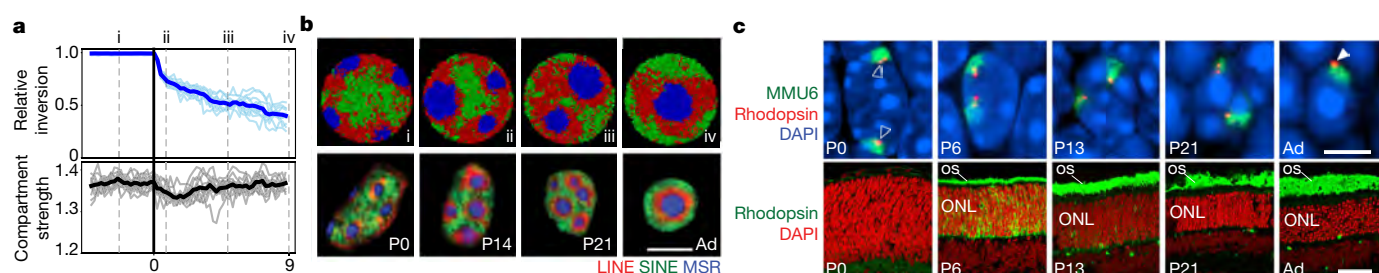


Fig. 4 | The time course and maintenance of compartment strength during nuclear inversion in the model and experiment. **a**, Simulated nuclear inversion. Configurations indicated by numerals and grey dashed lines are displayed in **b**. Black vertical line indicates the time at which interactions with the lamina are eliminated. Top, C monomers move towards the nuclear interior after removal of lamina interactions. Light-blue lines are computed from individual simulations; dark-blue line shows the average of the simulations. Bottom, compartment strength is maintained during inversion, showing only a transient dip. **b**, Representative conformations from simulations (top; see Extended Data Fig. 8a) mirror changes in chromatin architecture during rod differentiation in vivo at different developmental stages (postnatal day (P)0, P14, P21 and adult (Ad; 3.5 months); bottom) detected by FISH with probes for long interspersed nuclear elements (LINEs; L1, red),

short interspersed nuclear elements (SINEs; B1, green) and major satellite repeats (MSRs; blue). The progression of geometries remains unchanged when simulated inversion is accompanied by volume decrease (Extended Data Fig. 8c), in accordance with in vivo observations¹⁹. **c**, Top, in the process of nuclear inversion, the rhodopsin locus (red) within chromosome 6 (MMU6; green) changes position from internal (empty arrowheads) to peripheral (filled arrowhead) but remains within the A compartment (see Extended Data Fig. 9 for other genomic regions). Bottom, despite this marked relocation, rhodopsin gene expression, which starts at P6, continues at an increasing rate. OS, outer segments of rods positive for rhodopsin staining (green); ONL, outer nuclear layer containing rod perikarya. Single confocal sections (**b**, bottom) and projections of 2- μ m confocal stacks (**c**). Scale bars, 5 μ m (**b**, bottom and **c**, top) and 50 μ m (**c**, bottom).

- Solovei, I., Thanisch, K. & Feodorova, Y. How to rule the nucleus: *divide et impera*. *Curr. Opin. Cell Biol.* **40**, 47–59 (2016).
- van Steensel, B. & Belmont, A. S. Lamina-associated domains: links with chromosome architecture, heterochromatin, and gene repression. *Cell* **169**, 780–791 (2017).
- Bonev, B. & Cavalli, G. Organization and function of the 3D genome. *Nat. Rev. Genet.* **17**, 661–678 (2016).
- Jerabek, H. & Heermann, D. W. How chromatin looping and nuclear envelope attachment affect genome organization in eukaryotic cell nuclei. *Int. Rev. Cell Mol. Biol.* **307**, 351–381 (2014).
- Jost, D., Carrivain, P., Cavalli, G. & Vaillant, C. Modeling epigenome folding: formation and dynamics of topologically associated chromatin domains. *Nucleic Acids Res.* **42**, 9553–9561 (2014).
- Lee, S. S., Tashiro, S., Awazu, A. & Kobayashi, R. A new application of the phase-field method for understanding the mechanisms of nuclear architecture reorganization. *J. Math. Biol.* **74**, 333–354 (2017).
- Di Pierro, M., Zhang, B., Aiden, E. L., Wolynes, P. G. & Onuchic, J. N. Transferable model for chromosome architecture. *Proc. Natl Acad. Sci. USA* **113**, 12168–12173 (2016).
- Nuebler, J., Fudenberg, G., Imakaev, M., Abdennur, N. & Mirny, L. A. Chromatin organization by an interplay of loop extrusion and compartmental segregation. *Proc. Natl Acad. Sci. USA* **115**, E6697–E6706 (2018).
- van de Werken, H. J. G. et al. Small chromosomal regions position themselves autonomously according to their chromatin class. *Genome Res.* **27**, 922–933 (2017).
- Larson, A. G. et al. Liquid droplet formation by HP1 α suggests a role for phase separation in heterochromatin. *Nature* **547**, 236–240 (2017).
- Strom, A. R. et al. Phase separation drives heterochromatin domain formation. *Nature* **547**, 241–245 (2017).
- Machida, S. et al. Structural basis of heterochromatin formation by human HP1. *Mol. Cell* **69**, 385–397 (2018).
- Ganai, N., Sengupta, S. & Menon, G. I. Chromosome positioning from activity-based segregation. *Nucleic Acids Res.* **42**, 4145–4159 (2014).
- Grosberg, A. Y. & Joanny, J.-F. Nonequilibrium statistical mechanics of mixtures of particles in contact with different thermostats. *Phys. Rev. E* **92**, 032118 (2015).
- Smrek, J. & Kremer, K. Small activity differences drive phase separation in active-passive polymer mixtures. *Phys. Rev. Lett.* **118**, 098002 (2017).
- Stevens, T. J. et al. 3D structures of individual mammalian genomes studied by single-cell Hi-C. *Nature* **544**, 59–64 (2017).
- Hilbert, L. et al. Transcription organizes euchromatin similar to an active microemulsion. Preprint at <https://www.biorxiv.org/content/10.1101/234112v2> (2018).
- Zheng, X. et al. Lamins organize the global three-dimensional genome from the nuclear periphery. *Mol. Cell* **71**, 802–815 (2018).
- Solovei, I. et al. Nuclear architecture of rod photoreceptor cells adapts to vision in mammalian evolution. *Cell* **137**, 356–368 (2009).
- Solovei, I. et al. LBR and lamin A/C sequentially tether peripheral heterochromatin and inversely regulate differentiation. *Cell* **152**, 584–598 (2013).
- Eberhart, A. et al. Epigenetics of eu- and heterochromatin in inverted and conventional nuclei from mouse retina. *Chromosome Res.* **21**, 535–554 (2013).
- Tan, L., Xing, D., Chang, C.-H., Li, H. & Xie, X. S. Three-dimensional genome structures of single sensory neurons in mouse visual and olfactory systems. *Nat. Struct. Mol. Biol.* **26**, 297–307 (2019).
- Imakaev, M. et al. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat. Methods* **9**, 999–1003 (2012).
- Ou, H. D. et al. ChromEMT: visualizing 3D chromatin structure and compaction in interphase and mitotic cells. *Science* **357**, eaag0025 (2017).
- Choo, K. H. A. *The Centromere* (Oxford Univ. Press, 1997).
- Rubinstein, M. & Colby, R. H. *Polymer Physics* (Oxford Univ. Press, 2003).
- Biggs, R., Liu, P. Z., Stephens, A. D. & Marko, J. F. Effects of altering histone posttranslational modifications on mitotic chromosome structure and mechanics. *Mol. Biol. Cell* **30**, 820–827 (2019).
- Helminger, D. et al. Glutamine-expanded ataxin-7 alters TFCT/STAGA recruitment and chromatin structure leading to photoreceptor dysfunction. *PLoS Biol.* **4**, e67 (2006).
- Tang, S.-J. Chromatin organization by repetitive elements (CORE): a genomic principle for the higher-order structure of chromosomes. *Genes* **2**, 502–515 (2011).
- Rosa, A. & Everaers, R. Structure and dynamics of interphase chromosomes. *PLOS Comput. Biol.* **4**, e1000153 (2008).

Acknowledgements We thank S. Bultmann for help with rod cell sorting; A. S. Wang for help with sampling of *Lbr*^{-/-} thymic; D. Devys for samples of retinas from R7E mice; all members of the Mirny laboratory for many discussions; N. Abdennur for help with CTCF motif analysis; and N. Abdennur and P. Kerpedjiev for help with the HiGlass Hi-C browser. This work has been supported by NSF 1504942, NIH GM114190, NIH HG003143, NIH HG007743 and by the Deutsche Forschungsgemeinschaft grants SO1054/3 (I.S.) and SFB1064 (I.S. and H.L.). M.F. was supported by the Department of Defense through the National Defense Science & Engineering Graduate Fellowship (NDSEG) Program. J.D. and L.A.M. acknowledge support from the National Institutes of Health Common Fund 4D Nucleome Program (DK107980). J.D. is an investigator of the Howard Hughes Medical Institute.

Author contributions I.S., N.N., L.A.M. and J.D. conceived the project. Y.F. and I.S. obtained biological samples. N.N. performed Hi-C. M.F., M.I., G.F., B.R.L. and N.N. performed Hi-C analysis. M.F., with contributions from M.I., G.F. and L.A.M., performed simulations. Y.F., I.S. and B.J. conceived and performed microscopic experiments. M.F., Y.F. and I.S. prepared the figures. M.F., Y.F., G.F., I.S. and L.A.M. wrote the manuscript with contributions from N.N., M.I., H.L. and J.D.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-019-1275-3>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-019-1275-3>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to G.F., I.S. or L.A.M.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

METHODS

Data reporting. No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Cryosections and immunostaining. *Cryosections.* Retinas were sampled from CD1 mice at P0, P3, P6, P13, P21, P28 and at 3.5 months of age. Samples of retinas from R7E mice were provided by D. Devys. A detailed protocol of tissue fixation and cryosection preparation has been published previously³¹. In brief, tissues were fixed with 4% formaldehyde in PBS for 20–24 h, washed with PBS, incubated in solutions with increasing concentrations of sucrose (10%, 20% and 30%) and transferred into embedding moulds (Peel-A-Way Disposable Embedding Molds, Polysciences) filled with Jung freezing medium (Leica Microsystems). Tissue cryoblocks were frozen by immersing the moulds into a -80°C ethanol bath, and stored at -80°C . Cryosections with a thickness of 16–20 μm were cut using a Leica Cryostat (Leica Microsystems), collected on SuperFrost microscopic slides (SuperFrost Ultra Plus), immediately frozen and stored at -80°C before use.

Immunostaining. Rhodopsin expression during rod differentiation was studied with antibodies against rhodopsin (RET-P1, Abcam). Nuclear architecture of retinal cells in degenerating retina of R7E mice was studied using antibodies against the euchromatin marker of histone modification, H3K9ac (donated by H. Kimura), ATAXN7 (provided by D. Devys) and lamin A/C (provided by H. Herrmann). Secondary antibodies were conjugated to Alexa 488, Alexa 555, Alexa 594 or Alexa 647 (Invitrogen). A detailed description of the immunostaining protocol has been published previously³². In brief, sections were incubated with primary and secondary antibodies diluted in blocking solution (1% BSA, 0.1% Triton X-100 and 0.1% saponin) under glass chambers for 18–20 h at room temperature. Washes (3×30 min) in between and after antibody incubations were performed with 0.05% Triton X-100 in PBS at 37°C . For nuclear counterstaining, DAPI was added to the secondary antibody solution at a final concentration of 2 mg ml^{-1} .

FISH and microscopy. *FISH.* FISH on cryosections was performed according to the previously published protocol³¹. In brief, cryosections were dried for 30 min at room temperature, rehydrated in 10 mM sodium citrate buffer (pH 6.0) and heated in the same buffer for 30 min at 80°C for antigen retrieval. After equilibration with $2 \times \text{SSC}$ buffer and incubation with 50% formamide in $2 \times \text{SSC}$ for 30 min, probes were loaded onto cryosections under small glass chambers, sealed with rubber cement and pre-incubated on a heating block at 45°C for 1 h. Tissue and probe DNA were denatured simultaneously on a heating block at 80°C for 3–5 min. Hybridization was carried out at 37°C for 2 days. After hybridization, sections were washed with $2 \times \text{SSC}$ at 37°C and $0.1 \times \text{SSC}$ at 61°C , counterstained with $2 \mu\text{g ml}^{-1}$ DAPI for 1 h and mounted in Vectashield anti-fade medium (Vector Laboratories). *FISH probes.* BAC clones used in the study were purchased from BACPAC Resources (Children's Hospital Oakland). For coordinates of all BACs, see Supplementary Table 2. BAC DNA was amplified from a miniprep using GenomiPhi kit (GE Healthcare, UK), labelled by nick translation with fluorochrome-conjugated nucleotides and purified using QIAquick Nucleotide Removal Kit 50 (Qiagen). dUTPs were labelled with FITC, Cy3, Texas Red or Cy5 according to the published protocol³³. To verify BAC clones and exclude those that cross-hybridize to other chromosomes, all BAC probes were first labelled with digoxigenin–dUTP and co-hybridized with a respective chromosome paint labelled with biotin–dUTP to mouse metaphase spreads. Hybrids were detected with anti-digoxigenin antibody conjugated to FITC (Jackson Immuno Research) and avidin conjugated to Alexa 555 (Invitrogen Molecular Probes). Mouse chromosome paints were a gift from J. Wienberg. The paints were first amplified and then labelled with biotin–dUTP or Cy3–dUTP by degenerate oligonucleotide-primed polymerase chain reaction (DOP-PCR) using 6MW primer ($5'-\text{CCGACTCGAGNNNNNNATGTGG}-3'$, Eurogentec). For FISH probe preparation, 4 μg of labelled BAC or 6 μg of chromosome paint, were mixed with 10 μg of salmon sperm DNA and 50 μg of mouse Cot-1 DNA, ethanol-precipitated and dissolved in 10 μl of hybridization mixture consisting of 50% deionized formamide (Sigma-Aldrich), 10% dextran sulphate (Amersham Biosciences) and $1 \times \text{SSC}$ ³¹. Probes for FISH with SINEs (B1) and LINEs (LINE1) and major satellite repeats are described in a previous study¹⁹.

Immuno-FISH. For the nuclear lamina staining after FISH, sections were equilibrated in PBS and stained as described above using antibodies against lamin B1 (Santa Cruz, sc-6217) or lamin A/C and LBR (both provided by H. Herrmann).

Microscopy and image analysis. Image stacks were acquired using a Leica TCS SP5 confocal microscope equipped with Plan Apo 63 \times /1.4 NA oil-immersion objective and lasers for blue (405 nm), green (488 nm), orange (561 nm), red (594 nm) and far-red (633 nm) fluorescence. Multichannel image stacks were corrected for chromatic shift and processed using a dedicated ImageJ plugin 'Stack Groom'³⁴.

Mice. Mice used for tissue sampling were obtained from Charles River Laboratories, housed at the Biocentre, Ludwig Maximilians University of Munich (LMU) and treated according to the standard protocol approved by the Animal Ethics Committee of LMU.

Tissue sampling for Hi-C. Retinas from CD1 and C3H adult mice (retired breeders) were dissociated into a single-cell suspension using the Papain Dissociation System (Worthington Biochemical Corporation) as described elsewhere³⁵. Four retinas from two mice were used for one biological replica. To obtain a pure population of rod photoreceptors, retina suspensions were sorted based on standard forward and sideward scatter settings using FACS Aria II (Becton Dickinson) and yielded about 1 million rod perikarya (Supplementary Fig. 1). Retinas of C3H mice, which lack the entire outer nuclear layer, were used to obtain the non-rod population of retinal neurons. Each biological replica of non-rod neurons contained approximately 10 million cells. Thymocytes from wild-type CD1 mice and *Lbr*^{−/−} mice³⁶ were extracted from thymi of young adult animals at P26 and P28, respectively. Thymi were minced, small tissue pieces were gently pipetted and the resulting single-cell suspension was pressed through a Cell Strainer Snap Cap with a mesh size of 35 μm . Each biological replica of thymocytes contained 25 million–30 million cells. Images of microscopic controls of isolated rods, non-rod neurons and thymocytes are shown in Supplementary Fig. 1. All cells were fixed with 1% formaldehyde (Fisher Scientific, 10532955) for 10 min at room temperature. Fixation was quenched with 0.1 M glycine for 5 min at room temperature and then for 15 min at 4°C . Fixed cells were pelleted, snap-frozen and kept at -80°C until use.

Hi-C. Hi-C was performed as described previously³⁷ with modifications.

Cell lysis and chromatin digestion. In brief, 450,000 formaldehyde-cross-linked rod nuclei and up to 5 million of other cell types were incubated in 1 ml of cold lysis buffer (1 ml 10 mM Tris-HCl pH 8.0, 10 mM NaCl, 0.2% (v/v) Igepal CA630, mixed with 100 μl protease inhibitors (Sigma P8340) immediately before use) on ice for 15–20 min. Next, samples were lysed with a Dounce homogenizer and pestle A (KIMBLE Kontes, 885303-0002) by moving the pestle slowly up and down 25 times, incubating on ice for 1 min followed by 10 more strokes with the pestle. The suspension was centrifuged for 5 min at 9,800 r.p.m. (rod nuclei) and 4,500 r.p.m. (all other samples) at room temperature using a table-top centrifuge (Centrifuge 5810R, Eppendorf). The supernatant was carefully removed from the sample containing rod nuclei and spun a second time (9,800 r.p.m., 5 min) and then both pellets were combined. Pellets were washed twice with ice-cold 500 μl $1 \times \text{NEBuffer 2}$ (NEB). After the second wash, each pellet was resuspended in $1 \times \text{NEBuffer 2}$ in a total volume of 352 μl , chromatin was solubilized by addition of 38 μl 1% SDS per tube, the mixture was resuspended and incubated at 65°C for 10 min. Tubes were placed on ice and 44 μl of 10% Triton X-100 was added. Chromatin was subsequently digested by adding 400 units HindIII (NEB) at 37°C for 15 h with continuous slow rocking in parafilm-sealed tubes. Digested chromatin solutions were spun shortly, transferred to ice and used for generating Hi-C libraries.

Biotin marking of DNA ends and blunt-end ligation. The HindIII DNA ends were filled in and marked with biotin by adding 70 μl fill-in mix (2 μl 10 mM dATP, 2 μl 10 mM dGTP, 2 μl 10 mM dTTP, 42 μl 0.4 mM biotin-14-dCTP (Invitrogen 19518-018), 7 μl $10 \times \text{NEBuffer 2}$ and 15 μl 5 U μl^{-1} Klenow polymerase (NEB M0210L)) followed by incubation at 37°C for 4 h on a rocking platform at 50 r.p.m. Klenow polymerase was inactivated by adding 96 μl 10% SDS followed by incubation at 65°C for 30 min. Tubes were then immediately placed on ice; the content of each of the tube was transferred to a 15-ml conical tube containing 7.58 ml ligation mix (820 μl 10% Triton X-100, 758 μl $10 \times \text{ligation buffer}$ (500 mM Tris-HCl pH 7.5, 100 mM MgCl_2 and 100 mM DTT), 82 μl 10 mg ml^{-1} BSA, 82 μl 100 mM ATP and 5.84 ml water). Subsequently, 50 μl 1 U μl^{-1} T4 DNA ligase (Invitrogen, 15224) was added and mixed by inverting tubes; ligation was performed at 16°C overnight. For DNA purification, 50 μl of 10 mg ml^{-1} proteinase K (Invitrogen, 25530-031) was added to each tube and samples were incubated at 65°C for 4 h followed by a second addition of 50 μl of 10 mg ml^{-1} proteinase K solution and 8 h incubation at 65°C . Tubes were cooled to room temperature and DNA samples were transferred to 50-ml conical tubes. The DNA was extracted by adding an equal volume of phenol pH 8.0 (Fisher, BP1750I-400), vortexing for 3 min and spinning for 10 min at 4,000 r.p.m. in a table-top centrifuge (centrifuge 5810R, Eppendorf). The supernatants were transferred to new 50-ml conical tubes. Another two extractions were performed with an equal volume of phenol, pH 8.0:chloroform (1:1). Next, supernatants with Hi-C libraries were concentrated and desalted on 30-kDa Amicon Ultra 15-ml columns (Fisher Scientific, UFC903024) by spinning once for 10 min at 4,000 r.p.m. in a table-top centrifuge (centrifuge 5810R, Eppendorf). The flow-through was discarded and each column was washed once with 5 ml milliQ water. Then, samples were dissolved in 1 ml $1 \times \text{TE buffer}$, transferred to 30-kDa Amicon Ultra 0.5-ml columns (Fisher, UFC5030BK) and spun at 10,000 r.p.m. in a microcentrifuge. The flow-through was discarded. Columns were washed twice with 450 μl TE. After the final wash, the Hi-C library was dissolved in 100 μl water. Aliquots of Hi-C libraries were run on a gel to estimate the amount of DNA in the samples: 5 μl of rod Hi-C libraries and 2 μl for all other libraries.

Biotin removal from unligated ends. Hi-C libraries were treated with T4 DNA polymerase to remove biotinylated ends that did not ligate (dangling ends).

The reactions were assembled as follows: Hi-C library (up to 5 µg DNA), 1.3 µl 10 mg ml⁻¹ BSA, 13 µl 10× NEBuffer 2, 0.325 µl 10 mM dATP, 0.325 µl 10 mM dGTP and 30 units T4 DNA polymerase (NEB, M0203L) in a total volume of 130 µl. Reactions were mixed in a single tube, split between wells on a PCR plate and incubated at 20 °C for 5 h. Samples were pooled and the reaction was stopped by addition of 5.2 µl 0.5 M EDTA pH 8.0.

DNA fragmentation. The DNA was sheared to a size of 100–400 bp (with the majority of molecules around 200 bp) using a Covaris S2 instrument (Covaris). The settings were as follows: duty cycle 10%, intensity 5, cycles per burst 200, set mode frequency sweeping, process time 60 s per process, cycle number 3. DNA size was checked by running an aliquot on a 2.5% agarose gel and samples were sonicated for an additional half-cycle when deemed necessary, which allowed us to avoid library size selection. The DNA samples were purified using DNA MinElute columns (Qiagen, 5 µg DNA per column) and PB buffer (Qiagen). Elution was done in two steps with hot (65 °C) EB buffer so that the total volume of each Hi-C library was about 70 µl. DNA amounts were estimated to 5–9 µg of DNA per library by running aliquots on 2.5% agarose gel along with 100 ng of a low molecular weight DNA ladder (NEB, N3233L).

End repair and A' tailing. A single DNA end-repair reaction per Hi-C library was performed by adding 10 µl 10× ligation buffer (NEB, B0202S), 1.6 µl 25 mM dNTP mix, 5 µl T4 DNA polymerase (3 U µl⁻¹, NEB, M0203L), 5 µl T4 polynucleotide kinase (10 U µl⁻¹, NEB, M0201S), 1 µl Klenow DNA polymerase (5 U µl⁻¹, NEB, M0210S) and water up to 100 µl. The reaction was incubated at 20 °C for 1 h followed by purification of the DNA with a Qiagen MinElute column (Qiagen, up to 5 µg DNA per column). The DNA was eluted twice with 25 µl hot EB buffer (Qiagen). The eluates for each single column were pooled. Next, A'-tailing reactions that adenylate the 3' ends of the fragments were carried out by incubation with 7.5 µl 10× NEBuffer 2, 15 µl 1 mM dATP, 4.5 µl Klenow (exo-) (NEB, M0212L) and water to 75 µl. The reaction was incubated at 37 °C for 1 h followed by incubation at 65 °C for 20 min to inactivate the Klenow polymerase. The reactions were cooled on ice, all tubes for a library were pooled and the volume adjusted to 200 µl with 1× TLE buffer (10 mM Tris pH 8.0, 0.1 mM EDTA).

Streptavidin pull-down of biotinylated Hi-C ligation products. All subsequent steps were performed in DNA LoBind tubes (Eppendorf, 22431021) and each step was performed in a fresh tube. Then, 100 µl of streptavidin Dynabeads (MyOne Streptavidin C1 Beads, Invitrogen, 650-01) were washed twice with 400 µl Tween wash buffer (5 mM Tris-HCl pH 8.0, 0.5 mM EDTA, 1 M NaCl, 0.05% Tween-20) by incubating for 3 min at room temperature with rotation, reclaiming against a magnetic separation rack (Genscript, M00140) for 1 min and removing all supernatant. Next, reclaimed beads were resuspended in 200 µl 2× binding buffer (10 mM Tris-HCl pH 8.0, 1 mM EDTA, 2 M NaCl) and combined with 200 µl Hi-C DNA from the previous step. The mixture was incubated at room temperature for 30 min with rotation. The supernatant was removed and the DNA-bound streptavidin beads were washed once with 400 µl 1× binding buffer. The beads were then washed with 100 µl 1× ligation buffer (Invitrogen, 5× buffer) with extra ATP (4 µM final concentration), and then resuspended in 38.8 µl 1× ligation buffer.

Paired-end adaptor ligation. Ligation reactions were prepared as follows: 38.8 µl Hi-C library on beads, 6 µl Illumina paired-end adapters (Illumina), 2.25 µl 5× ligation buffer (Invitrogen, supplied with T4 DNA ligase), 3 µl T4 DNA ligase (Invitrogen, 15224). The reaction was incubated at room temperature for 5 h. The beads with bound ligated Hi-C DNA were collected by holding the tubes against a magnetic separation rack (Genscript, M00140), washed twice with 400 µl 1× Tween wash buffer for 5 min on a rocking platform, once with 200 µl 1× binding buffer, twice with 200 µl 1× NEBuffer 2 to remove non-ligated paired-end adapters and resuspended in 18 µl 1× NEBuffer 2.

Library amplification. Six PCR reactions per library were set up, each containing 3 µl Dynabead-bound Hi-C library, Illumina PE1.0 and PE2.0 PCR primers (0.7 µl each; corresponding to 17.5 pmol each), 0.4 µl 25 mM dNTPs, 1 µl Pfu Ultra II Fusion DNA polymerase (Stratagene, 600670), 5 µl 10× Pfu Ultra buffer and 39.2 µl water. The temperature profile during the PCR amplification was 30 s at 98 °C followed by nine cycles of 10 s at 98 °C, 45 s at 65 °C, 30 s at 72 °C and a final 7-min extension at 72 °C. The PCR reactions were pulled together, streptavidin beads were collected using a magnetic separation rack for 2 min and supernatants were transferred to new tubes. Hi-C libraries were purified from the supernatants using Ampure XP beads (Beckman Coulter, A63881) as follows: 1.8× volumes of the beads were added to Hi-C samples, briefly vortexed, incubated at room temperature for 10 min and then collected using a magnetic rack for 5 min. Supernatants were discarded and beads were washed twice with 1 ml freshly made 70% ethanol. Air-dried beads were resuspended in 35 µl TLE buffer, incubated at room temperature for 15 min while tapping the tubes every 1–2 min and collected using a magnetic rack for 5 min. Supernatants were transferred to fresh tubes. The quality of Hi-C libraries was confirmed by NheI restriction digest of 8 µl of each Hi-C library. Digested samples were run in parallel with undigested samples on a 2% agarose gel. More than 50% of each Hi-C library was digested,

and all libraries were qualified for sequencing on an Illumina GAII paired-end sequencing platform.

Data analysis for Hi-C. *Hi-C data processing.* We mapped our reads to the mm9 genome assembly, and subsequently filtered and corrected the reads using ICE as described previously²³. We removed bins with less than half of the bin sequenced, in addition to bins at the lowest 1% of coverage. We truncated the top 0.05% of *trans* contacts, which were probably PCR blowouts. Read statistics can be found in Supplementary Table 1, with a comparison to other primary tissue datasets.

Compartment profile. In order to define compartment strength, it is necessary to have a particular assignment of Hi-C bins to compartments. For simulations, we know the sequence of A and B monomers along our simulated chromosomes. Hence, we can make the choice that a bin in our simulated Hi-C map is an A- or B-compartment bin if the majority of monomers belonging to that bin are A or B monomers. For experimental data, the process is more involved. For each chromosome, we take the *cis*-contact map, and following iterative correction and removal of distance decay to produce an 'observed over expected' matrix²³, we compute eigenvectors of the mean-centred observed-over-expected matrix. The eigenvector with the largest magnitude eigenvalue is the 'compartment signal'. However, the mathematics of this operation leaves the sign of the eigenvector ambiguous, although the partitioning of the genome into two separate compartments suggests it is not. The established convention is that the sign of this eigenvector is chosen such that the compartment signal correlates positively with GC content²³ or density of transcription start sites³⁸. In this convention, B-compartment bins are those for which the compartment signal is negative, and A compartment bins are those for which the compartment signal is positive.

Saddle plots. For each chromosome, we sort the compartment eigenvector from the lowest to the highest value. We then reshuffle the observed-over-expected map of the chromosome according to this ordering. We coarse-grain the resulting map into a 50-by-50 matrix, where the element (*i*, *j*) is the average value in the reshuffled map between bins of the *i*th 50-cile and the *j*th 50-cile. The saddle plot is the average of these coarse-grained maps over all chromosomes in both replicates. Analysis was performed at a resolution of 50 kb per bin.

Compartment strength. Given an assignment of bins to compartments, we define compartment strength first on a per-bin level. The compartment strength of bin *i* (*CS_i*) is the average number of contacts it makes with other bins of the same compartment type in the observed-over-expected heat map, divided by the average number of contacts it makes with any bin in the observed-over-expected heat map. The compartment strength of the total data set is then <*CS_i*>, where the average is taken over all bins, weighted equally. Note that this metric is independent of the orientation of the compartment profile, since the two compartments are treated symmetrically. If there is no compartmentalization, the metric is 1, whereas any pattern of compartmentalization yields a compartment strength greater than 1.

TAD strength. Two methods were used to analyse TAD strength. First, on the basis of the calls from a previous study³⁹, each TAD was rescaled such that it was a 30-by-30 bin heat map, and averaged together with other TADs within the same chromosome (Extended Data Fig. 3b). For each of these rescaled TADs, we computed their observed-over-expected maps, and compared the sum of their corners to the average of the two triangles adjacent to the corner. The side of each triangle was 12 bins. This is illustrated in the schematic shown in Extended Data Fig. 3b. TAD strength was then computed as the average of these values.

Second, TADs were called using corner score, implemented in the package lavaburst (<https://github.com/nvictus/lavaburst>) with default parameters (Extended Data Fig. 3f). Average enrichments of TADs were then calculated as described previously⁴⁰. For each TAD call, we took a matrix that was three times the size of a TAD, with a TAD located in the centre of the matrix. The matrix was then rescaled to a 90 × 90 matrix, with the TAD occupying the central (30:60, 30:60) square. The average TAD was obtained by averaging these 90 × 90 matrices. TAD strength was also calculated similarly to a previous study⁴⁰. It was defined as a ratio (2× within TAD)/(between TAD), where 'within TAD' is the sum of counts inside the TAD (30:60, 30:60) in the rescaled 90 × 90 matrix. 'Between TAD' is a sum of the counts between the TAD, and the regions before and after of the same length: (0:30, 30:60) and (30:60, 60:90) in the 90 × 90 matrix.

Insulation profiles. Insulation profiles are calculated following a previously published method⁴¹, removing two diagonals from each side of the main diagonal. Loci within two bins of a bad bin were also excluded. A window of size 200 kb was used with data at a resolution of 20 kb.

cis contact fraction. To quantify the territoriality of our data, we divided the number of *cis* (same chromosome, greater than 20 kb apart) reads by the sum of *cis* and *trans* (different chromosome) reads.

P(s) curves. The decay of contact probability as a function of distance from the diagonal is computed at the fragment level. *P(s)* curves were normalized such that they cross at 1 Mb.

Simulations. We perform Langevin dynamics of our coarse-grained model using a laboratory-developed wrapper for OpenMM^{42,43}, a high-performance

GPU-assisted MD API. Our chromosomes are constructed from equally sized spherical monomers with diameters defined to be of unit length. A rough estimate for how many base pairs each monomer represents is based on the identification of each polymer with one mouse chromosome.

Our simulations represent multiple copies of mouse chromosomes 1 and 2. The first 1,000 monomers proximal to the centromere region of each chromosome were assigned to be C monomers. The subsequent 5,000 monomers were A and B monomers mimicking the assignment of compartments in chromosome 1. We digitized the compartment eigenvector of chromosome 1, binned at 200 kb, and assigned five monomers to each of the first 1,000 bins to be A or B monomers if the corresponding eigenvector value was positive or negative, respectively. Chromosome 2 was represented similarly in simulations, but starting with the eigenvector of chromosome 2. To improve averaging of simulation observable values, our full system consisted of four copies of the chromosome 1- and four copies of the chromosome 2-derived sequences. Each monomer therefore represents 40 kb.

Unless otherwise noted, polymers were initialized as random walks. Preliminary simulations to determine orderings of parameter strengths were run for 2×10^6 time steps. Conventional parameter sweep simulations were run for 1.1×10^7 time steps and inverted parameter sweep simulations were run for 2.1×10^7 time steps, to allow for equilibration of compartment strength. Inversion simulations were initialized as the final configurations of conventional nuclei simulations, and were run for 0.9×10^7 time steps, with removal of the lamina occurring a quarter of the way through, after 2.25×10^6 time steps. Simulations of alternative models were run for 4.5×10^6 time steps, and de-inversion simulations were run for 2×10^7 .

We used six different energies in our equilibrium simulations: a stretching energy between pairs of adjacent monomers, a harmonic bending energy for triplets of monomers, spherical confinement, short-range attraction of B and C monomers to the lamina, a short-range inter-monomer attraction of varying strength, and a pinning of C monomers to the lamina. Details and functional forms can be found in Supplementary Information 1.

Simulated Hi-C heat maps were generated by counting contacts between pairs of loci over multiple simulation snapshots from multiple simulations. A contact was registered if the centres of two monomers were closer than 2.5 monomer diameters. For both the inverted and conventional model parameter sweeps, each data point represented contacts from the final 125 configurations of three separate simulations, with each configuration separated by 3,000 time steps. For enrichments over the inversion process, each data point was calculated from contacts obtained from 60 configurations drawn from eight separate simulations. For comparisons with Hi-C, after tallying contacts for the full simulation, any corresponding contacts that respond to contacts with C monomers were removed, as those represent regions that are not assayed in Hi-C due to low mappability. The resulting simulated Hi-C heat maps were then iteratively corrected and compartment strengths were computed in the same way as for experimental data.

For particular points in parameter space for which we wanted to display simulated Hi-C maps (Fig. 3b, e), 250 configurations from 50 simulations (for a total of 12,500 configurations) were necessary to smoothly sample the entire map.

Simulated configurations were compared quantitatively to microscopy through the distributions of each monomer type as a function of nuclear radius. For each monomer, we calculated its radial distance, normalized by the radius of the nucleus, and then binned according to the binning used previously¹⁹. Thus, for any configuration or group of configurations, we produced three distributions of monomer density as a function of nuclear radius, one for each of the three monomer types. For each distribution calculated in this way, we identified the radial distance at which the distribution achieved its maximum. We then computed the Euclidean distance between the peaks of our models and the peaks of the density functions described previously¹⁹, to quantitatively compare the performance of our models with respect to microscopy data. In figures, we refer to this metric as the density peak distance.

To ensure that our results are not sensitive to the choice of metric, we compare our density peak distance to two other measures of probability distribution function distance (Supplementary Fig. 4). These include Kullback–Leibler divergence (with reference distribution being the experimentally determined distribution), and the L^2 norm of the difference between the two distributions. Specifically, for each model class, and each monomer type, we compute the radial distributions, and then concatenate the three monomer-type distributions together. These are then compared (either with the L^2 norm or Kullback–Leibler divergence) to the similarly concatenated experimentally determined distribution. Good agreement with experiment was defined as being below the minimum value of density peak distance achieved in the parameter sweep plus $1.6 \times \text{s.d.}$ at that minimum value point.

Various geometrical aspects of the inversion process were also quantified. In Fig. 4, we track the average distance of the chromocentres from the nuclear centre, and normalize by the radius of the nucleus. In Extended Data Fig. 8, we track the average pairwise distance of all the chromocentres, normalized by the maximum pairwise distance. In both figures, we show the individual traces, computed from just one configuration, and then the average of the traces over 10 replicate

simulations. For Extended Data Fig. 8c, we increased the density from 0.15 to 0.55 in increments of 0.02, restarting our simulation every 225,000 time steps.

Choosing parameters for model-space exploration. To explore the six-dimensional space of our copolymer framework, we selected six energies and permuted them in terms of their assignments to the six possible attractions (Fig. 2a). The energies we chose were 0.02, 0.10, 0.20, 0.26, 0.34, and 0.44 (in units of kT). We selected these values such that for a sequence XX, XY, YY: $XY < (XX + YY)/2$, thereby satisfying the Flory–Huggins criterion for demixing of XX and YY. Thus, we expect that for any model class (which we define as a particular ordering of the attractions) the phase separation between XX and YY can take place.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

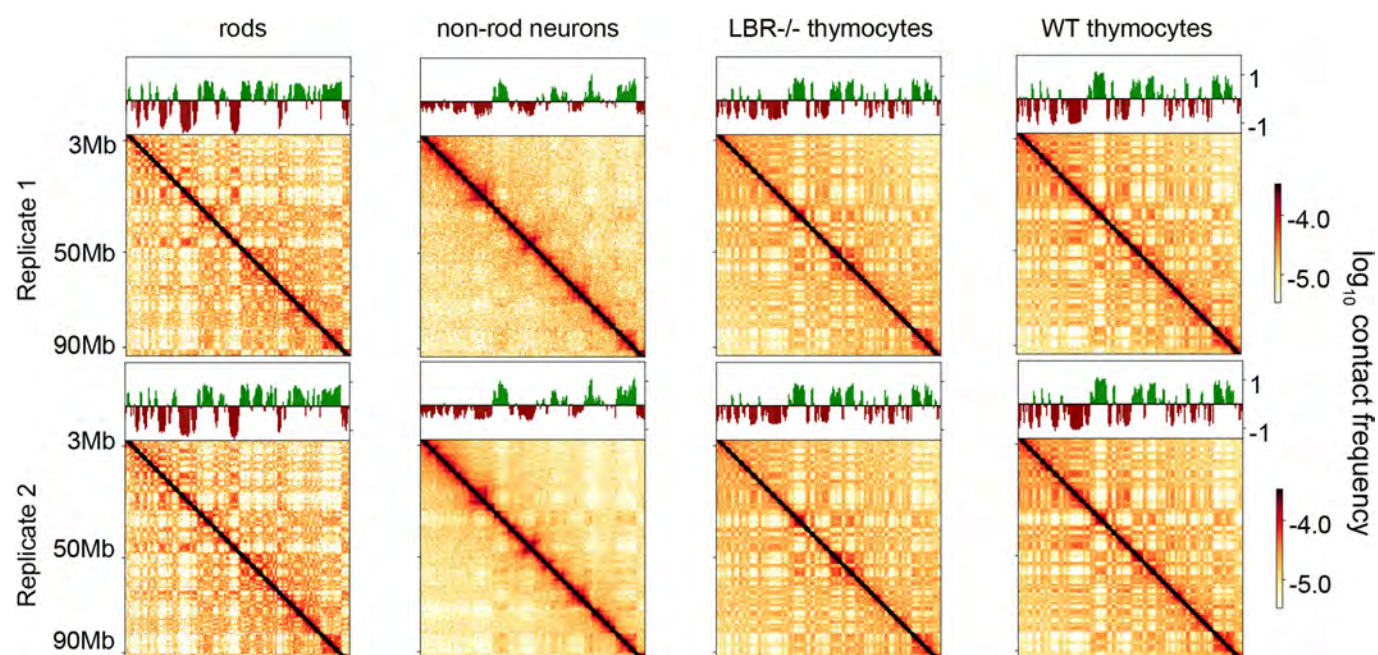
Data availability

Hi-C maps are available from the HiGlass browser (<http://mirnylab.mit.edu/projects/invnuclei/>) and from a public server (<http://higlass.io/app/?config=JLOhiPILtmq6qDRicHmJqg>). Hi-C maps are also available from the Gene Expression Omnibus (GEO) repository, accession number GSE111032.

Code availability

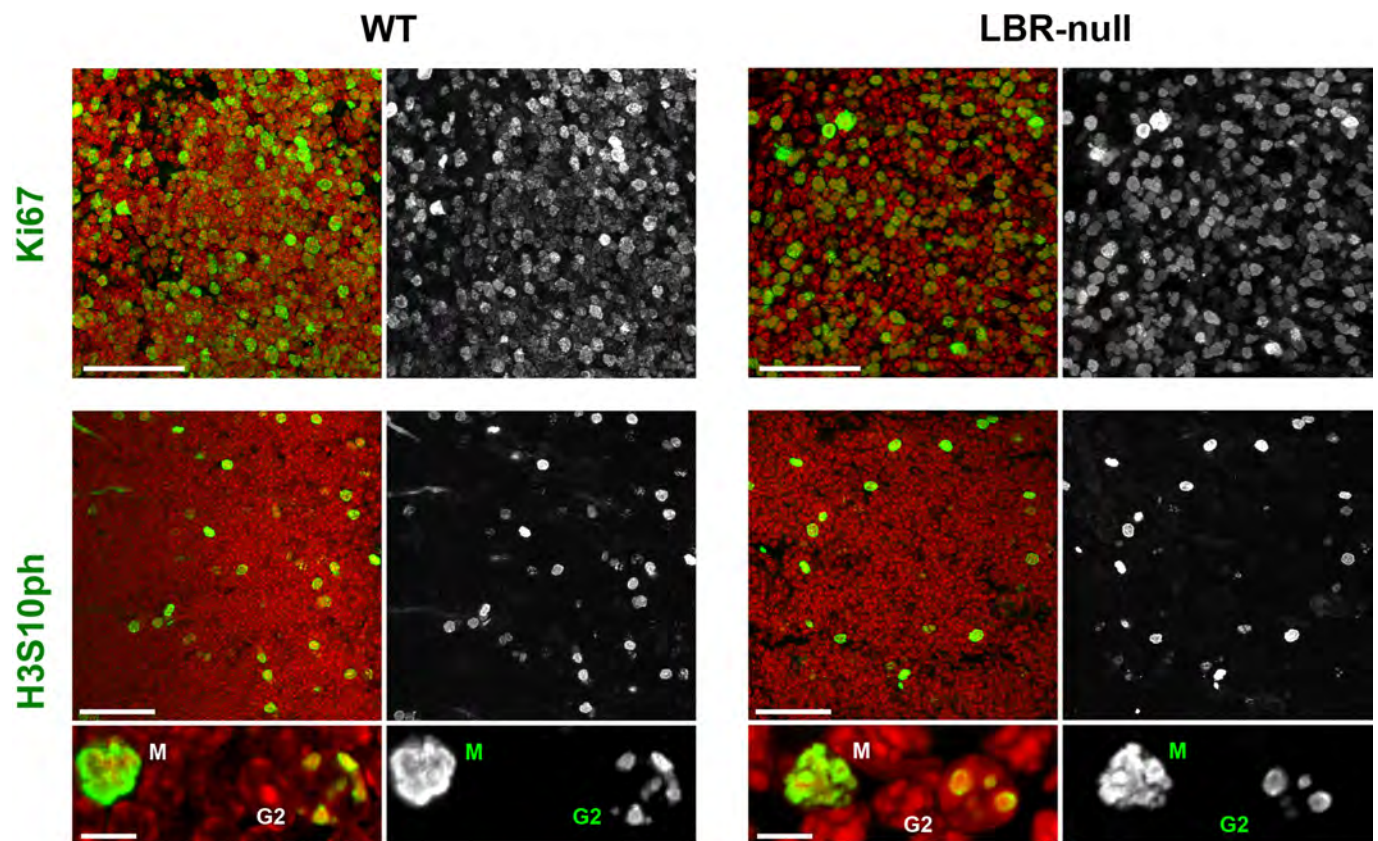
Software used to store and analyse Hi-C data can be accessed at <https://bitbucket.org/mirnylab/hiclib> and <https://bitbucket.org/mirnylab/mirnylib>. Data were also stored using the Cooler⁴⁴ software (<https://github.com/mirnylab/cooler>).

- Solovei, I. Fluorescence in situ hybridization (FISH) on tissue cryosections. in *Methods Mol. Biol.* **659**, 71–82 (2010).
- Eberhart, A., Kimura, H., Leonhardt, H., Joffe, B. & Solovei, I. Reliable detection of epigenetic histone marks and nuclear proteins in tissue cryosections. *Chromosome Res.* **20**, 849–858 (2012).
- Cremer, M. et al. Multicolor 3D fluorescence in situ hybridization for imaging interphase chromosomes. *Methods Mol. Biol.* **463**, 205–239 (2018).
- Walter, J. et al. Towards many colors in FISH on 3D-preserved interphase nuclei. *Cytogenet. Genome Res.* **114**, 367–378 (2006).
- Feodorova, Y., Koch, M., Bultman, S., Michalak, S. & Solovei, I. Quick and reliable method for retina dissociation and separation of rod photoreceptor perikarya from adult mice. *Methods* **2**, 39–46 (2015).
- Cohen, T. V. et al. The lamin B receptor under transcriptional control of C/EBP ϵ is required for morphological but not functional maturation of neutrophils. *Hum. Mol. Genet.* **17**, 2921–2933 (2008).
- Naumova, N. et al. Organization of the mitotic chromosome. *Science* **342**, 948–953 (2013).
- Heinz, S. et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010).
- Nora, E. P. et al. Targeted degradation of CTCF decouples local insulation of chromosome domains from genomic compartmentalization. *Cell* **169**, 930–944 (2017).
- Flyamer, I. M. et al. Single-nucleus Hi-C reveals unique chromatin reorganization at oocyte-to-zygote transition. *Nature* **544**, 110–114 (2017).
- Crane, E. et al. Condensin-driven remodelling of X chromosome topology during dosage compensation. *Nature* **523**, 240–244 (2015).
- Eastman, P. et al. OpenMM 4: a reusable, extensible, hardware independent library for high performance molecular simulation. *J. Chem. Theory Comput.* **9**, 461–469 (2013).
- Eastman, P. et al. OpenMM 7: rapid development of high performance algorithms for molecular dynamics. *PLoS Comput. Biol.* **13**, e1005659 (2017).
- Abdennur, N. & Mirny, L. Cooler: scalable storage for Hi-C data and other genomically-labeled arrays. Preprint at <https://www.biorxiv.org/content/10.1101/557660v1> (2019).
- Shultz, L. D. et al. Mutations at the mouse ichthyosis locus are within the lamin B receptor gene: a single gene model for human Pelger–Huët anomaly. *Hum. Mol. Genet.* **12**, 61–69 (2003).
- Dixon, J. R. et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380 (2012).
- Rao, S. S. P. et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).
- Schwarzer, W. et al. Two independent modes of chromatin organization revealed by cohesin removal. *Nature* **551**, 51–56 (2017).
- Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017–1018 (2011).
- Schmidt, D. et al. Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. *Cell* **148**, 335–348 (2012).
- Kerpedjiev, P. et al. HiGlass: web-based visual exploration and analysis of genome interaction maps. *Genome Biol.* **19**, 125 (2018).
- Zhang, Y. et al. Spatial organization of the mouse genome and its role in recurrent chromosomal translocations. *Cell* **148**, 908–921 (2012).
- Lin, Y. C. et al. Global changes in the nuclear positioning of genes and intra- and interdomain genomic interactions that orchestrate B cell fate. *Nat. Immunol.* **13**, 1196–1204 (2012).
- Kizilyaprak, C., Spehner, D., Devys, D. & Schultz, P. In vivo chromatin organization of mouse rod photoreceptors correlates with histone modifications. *PLoS ONE* **5**, e11039 (2010).



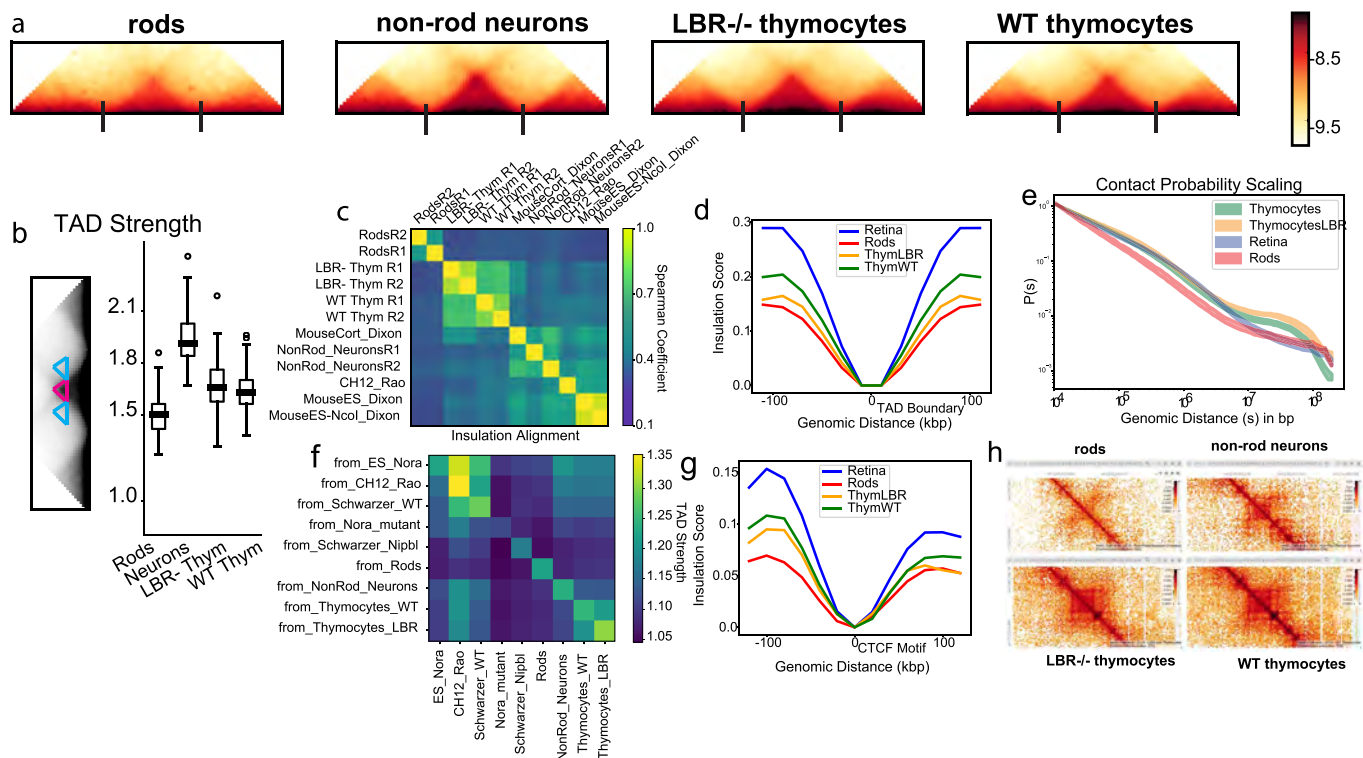
Extended Data Fig. 1 | Hi-C replicates show reproducible features. Hi-C maps are qualitatively similar between replicates. Hi-C maps (plotted as $\log_{10}(\text{contact frequencies})$) for an 87-Mb region of chromosome 1. Compartment profiles indicating regions in the A (green) and B

(red-brown) compartments are shown above the Hi-C maps. Full maps are available to browse on the HiGlass website (<http://higlass.io/app/?config=JLOhiPILTmq6qDRicHMJqg>). For quantitative comparisons, see Extended Data Figs. 3, 4, 5.



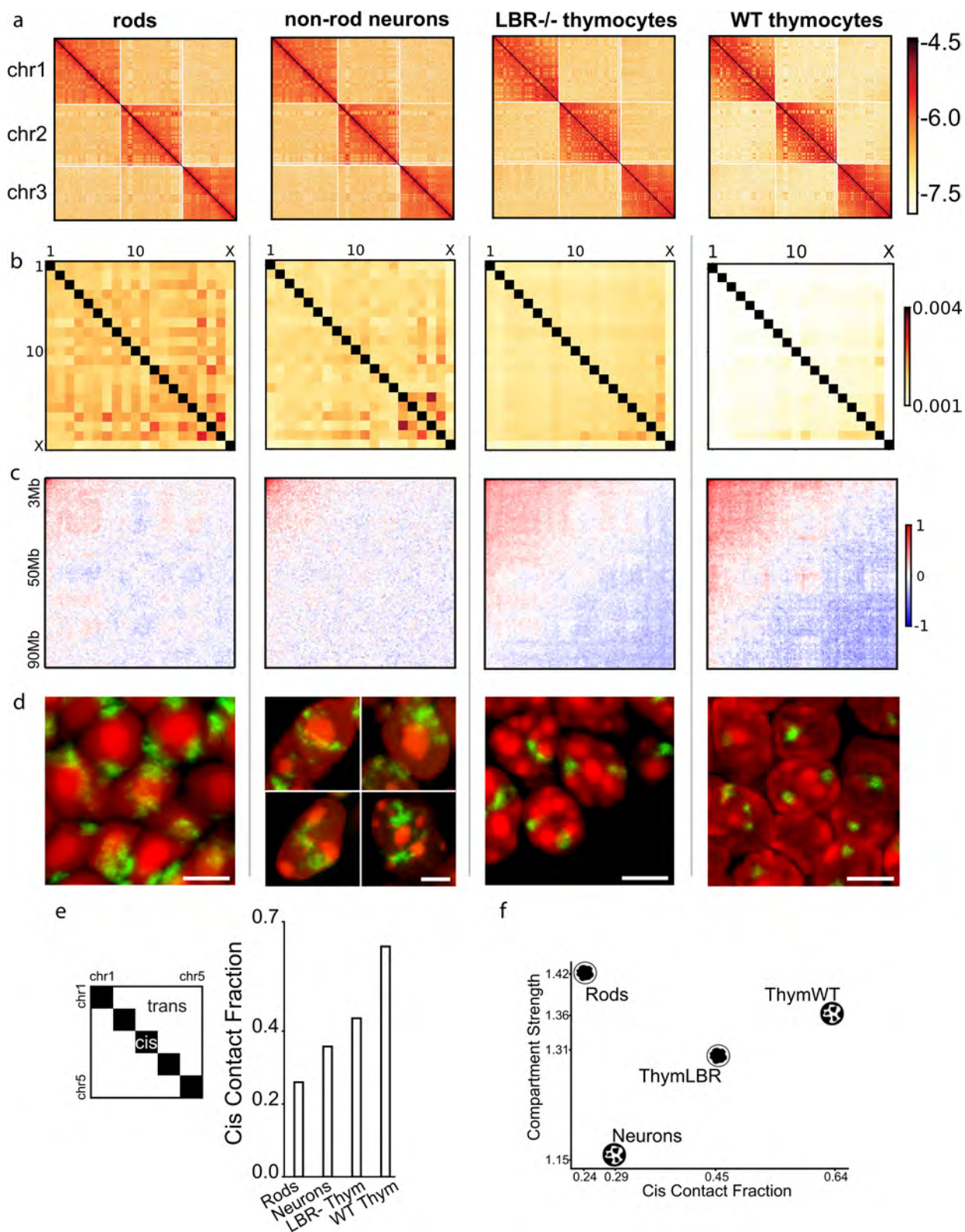
Extended Data Fig. 2 | The majority of thymocytes are actively cycling cells in both wild-type and *Lbr*^{-/-} mice. Left, wild-type mice; right, *Lbr*^{-/-} mice. Thymus cryosections were immunostained with antibodies for Ki-67, a marker of cycling cells, and phosphorylated histone H3 S10 (H3S10ph), a marker for G2 and mitotic cells. In agreement with the idea that *Lbr*^{-/-} mice have a seemingly normal immune system⁴⁵, the number of cycling thymocytes in thymi of *Lbr*^{-/-} mice is comparable to

that of wild-type mice. M, mitotic cells; G2, cells in mid/late G2. Ki-67 staining is shown as projections of 5- μ m confocal stacks. Phosphorylated H3 S10 staining is shown as projections of 10- μ m (for overviews) or 3- μ m (for magnified areas) confocal stacks. Antibodies: mouse anti-phosphorylated H3 S10 (Abcam, ab14955) and rabbit anti-Ki-67 (Abcam, ab15580). Immunostaining and microscopy were performed as described in the Methods. Scale bars, 50 μ m (top and middle) and 5 μ m (bottom).



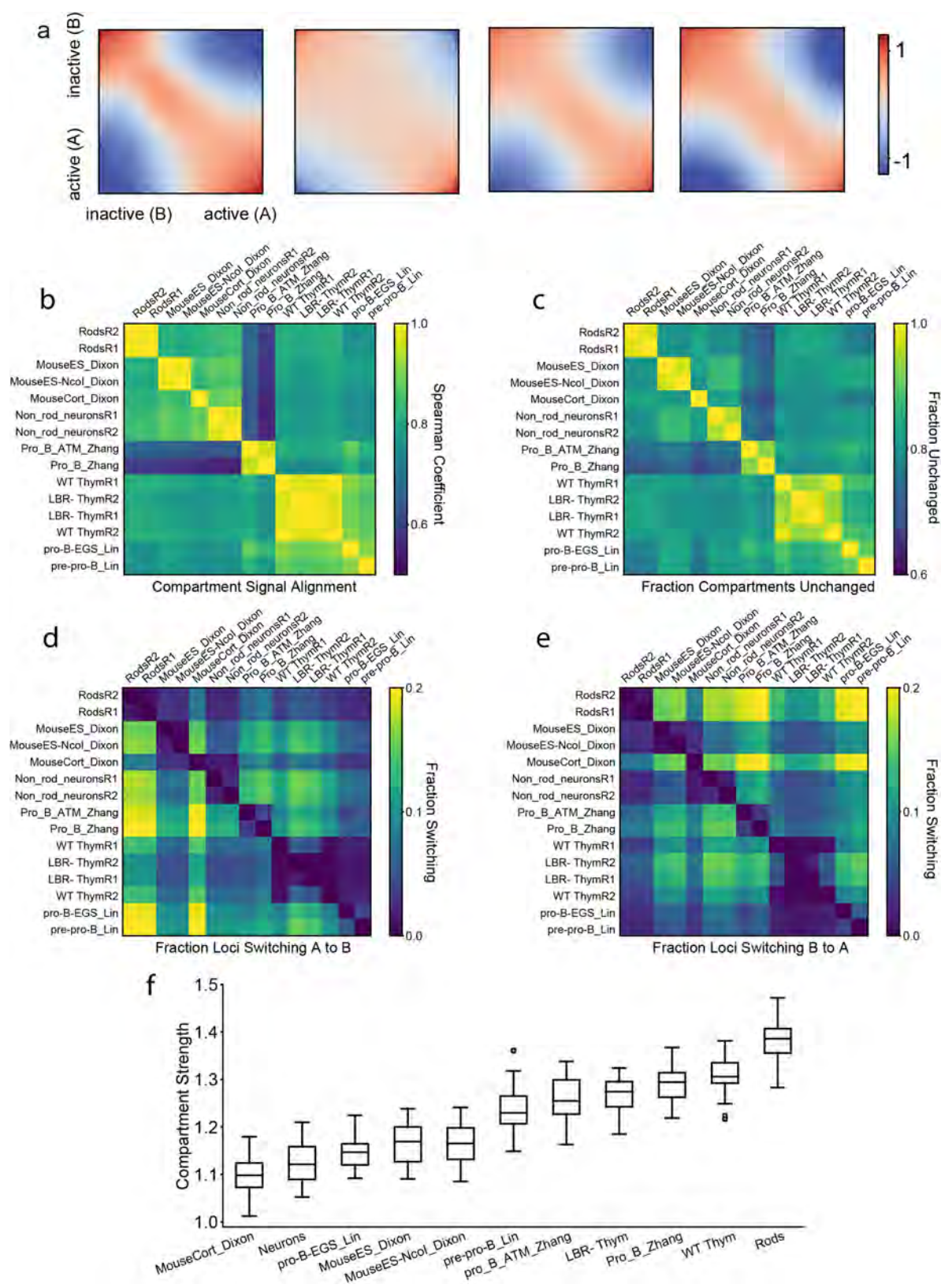
Extended Data Fig. 3 | Quantitative analysis of TADs. **a**, Average TADs, based on domain calls from embryonic stem cells³⁹. Ticks indicate start and end of TADs. The visual suggestion is that TADs are weakest in rods and strongest in non-rod neurons, with wild-type and *Lbr*^{-/-} thymocytes having intermediate strength. **b**, TAD strength is weakest in rods and strongest in non-rod neurons. TAD strength is the ratio of average contacts within the TAD (pink triangle) to average contacts between TADs (blue triangles). TAD strength is calculated separately for each autosome in two replicates. $n = 38$ chromosomes. Centre line is the median, the box ranges from the lower to upper quartiles and whiskers extend to $1.5 \times$ the interquartile range. **c**, Spearman correlation of insulation profiles across multiple mouse cell types, clustered hierarchically. Data were obtained from previous studies (GEO accession numbers GSE35156 and GSE63525)^{46,47}, as indicated by the name of the first author in the row and column labels. ES, embryonic stem cells. **d**, Average insulation profile (Methods) around TAD boundaries called in embryonic stem cells³⁹. The minimum insulation score of each profile is set to zero. We symmetrize noise by reflecting around the TAD boundary and averaging the reflected and original profiles. **e**, Decay of contact probability, $P(s)$, as a function of genomic separation, s . Shaded areas are bounded by $P(s)$ curves for biological replicates. All $P(s)$ curves are normalized to

their value at 10 kb. For rods, the steeper slope below 1 Mb and lack of a rollover in contrast to the other three cell types is indicative of weaker TADs, as previously described⁴⁸. **f**, TAD strength as a function of cell type (columns) and cell type from which TADs are called (rows). Data were obtained from previous studies (GEO accession numbers GSE98671, GSE63525 and GSE93431)^{39,47,48}, as indicated by the name of the first author in the row and column labels. Note that rods cluster with cell types with demonstrated weaker TADs. TAD strength is computed with the lavaburst approach (see Methods). **g**, Average insulation profile (Methods) oriented around the top 10^4 scoring CTCF motifs. For scoring, we used the FIMO algorithm⁴⁹, with a position weight matrix for the M1 motif as previously described⁵⁰. The minimum insulation score of each profile is set to zero, and the CTCF motif points to the left. This provides a TAD-call independent method of inferring TAD strength, given that CTCF is frequently present at the borders of TADs. **h**, Snapshot of HiGlass⁵¹ view of the four datasets, close to the diagonal (chromosome 12: 77,538,523–85,180,785 and chromosome 12: 79,240,367–82,837,977; 32-kb resolution). Rods almost completely lack TADs and non-rod neurons have very strong TADs, upon inspection. Datasets can be browsed in a more in-depth fashion on the HiGlass website (<http://higlass.io/app/?config=JLOhiPILtmq6qDRicHMJqg>).



Extended Data Fig. 4 | Quantitative analysis of territories. **a**, Hi-C contact maps for chromosomes 1, 2 and 3 show both a checkerboard pattern in *cis* (within a chromosome) and *trans* (between chromosomes), reflecting compartmentalization, and more frequent *cis* than *trans* contacts, reflecting chromosome territoriality. Views are shown for the second biological replicate, binned at 500 kb. **b**, Average number of contacts between pairs of chromosomes. Average *cis* contacts are much higher than *trans* contacts. Maps are normalized by their sums. **c**, Average contacts *in trans*. For every unique pair of chromosomes, we averaged the first 60 Mb, binned at 500-kb resolution. Maps are normalized to their means and plotted in log-space. There is evidence of weak enrichment

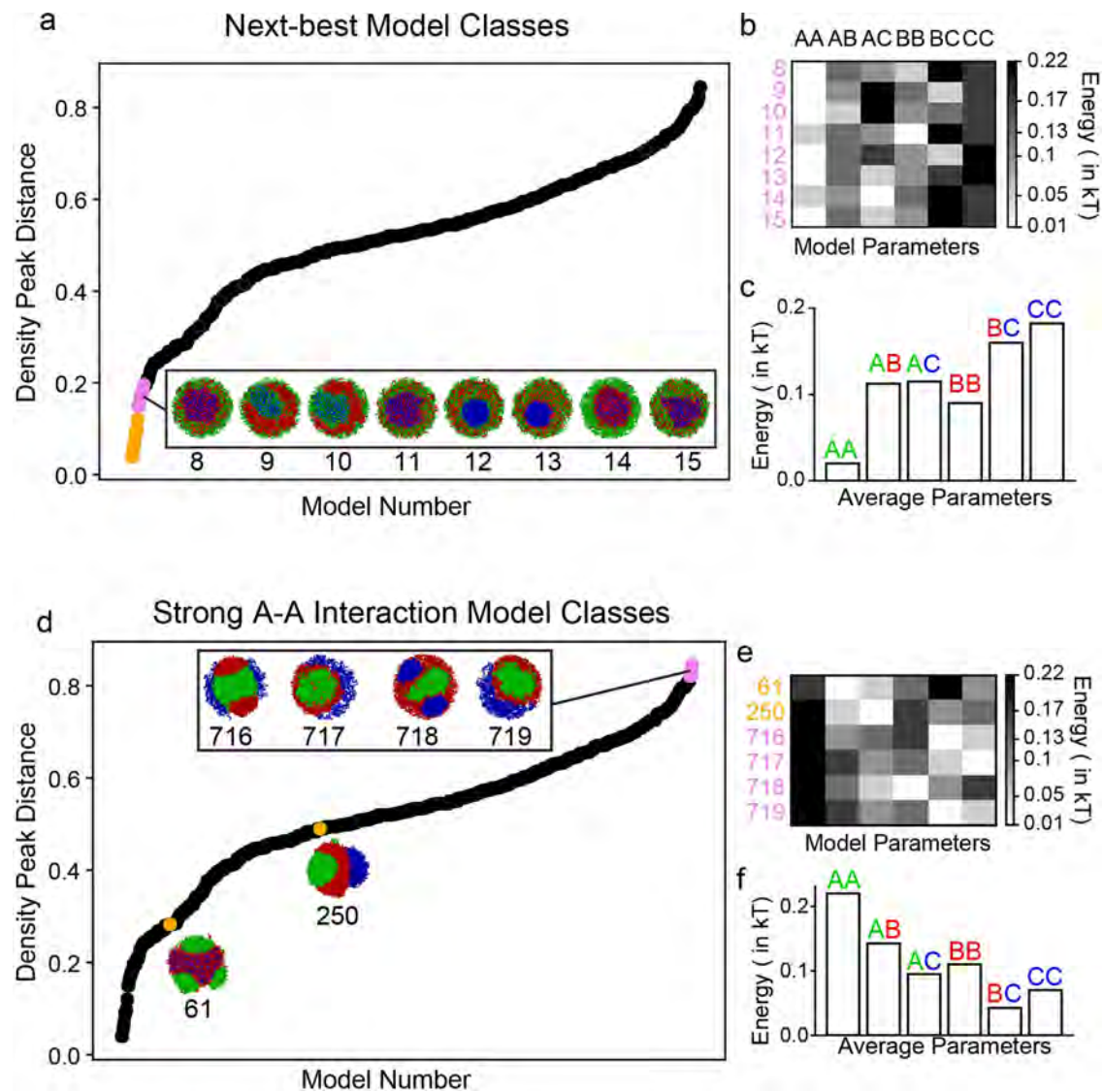
among chromocentre-proximal regions *in trans*, independent of inversion status. **d**, Consistent with the low *cis*-contact fraction revealed by Hi-C, chromosome 11 visualized by FISH (green) has a more diffuse territory in post-mitotic rods and non-rod neurons in comparison to cycling thymocytes of both genotypes. Projections of 2- μ m confocal stacks. Scale bars, 5 μ m. The chromosome painting was performed in four independent experiments. **e**, Chromosome territoriality, measured as the ratio of *cis* contacts to *cis* and *trans* contacts, is weaker in rods and non-rod neurons in comparison to conventional and inverted thymocytes. The schematic illustrates the compared regions. **f**, Scatterplot of compartmentalization and territoriality. The two metrics are not necessarily related.



Extended Data Fig. 5 | See next page for caption.

Extended Data Fig. 5 | Quantitative analysis of compartments. **a**, Saddle plots²³ (see Methods) of contact frequency enrichment show the extent of compartmentalization across cell types *in cis*. **b**, Spearman correlation of compartment profiles across multiple mouse cell types, clustered hierarchically. Data were obtained from previous studies (GEO accession numbers GSE35156, GSE35519 and GSE40173)^{46,52,53}, as indicated by the name of the first author in the row and column labels. Spearman's $r(\text{LBR1, WT1}) = 0.95$, $P < 1 \times 10^{-10}$, $n = 4,780$; $r(\text{LBR1, LBR2}) = 0.98$, $P < 1 \times 10^{-10}$, $n = 4,780$; $r(\text{WT1, WT2}) = 0.99$, $P < 1 \times 10^{-10}$, $n = 4,780$. P values are from two-sided tests. Positions of compartments are almost exactly the same between wild-type thymocytes and *Lbr*^{-/-} thymocytes,

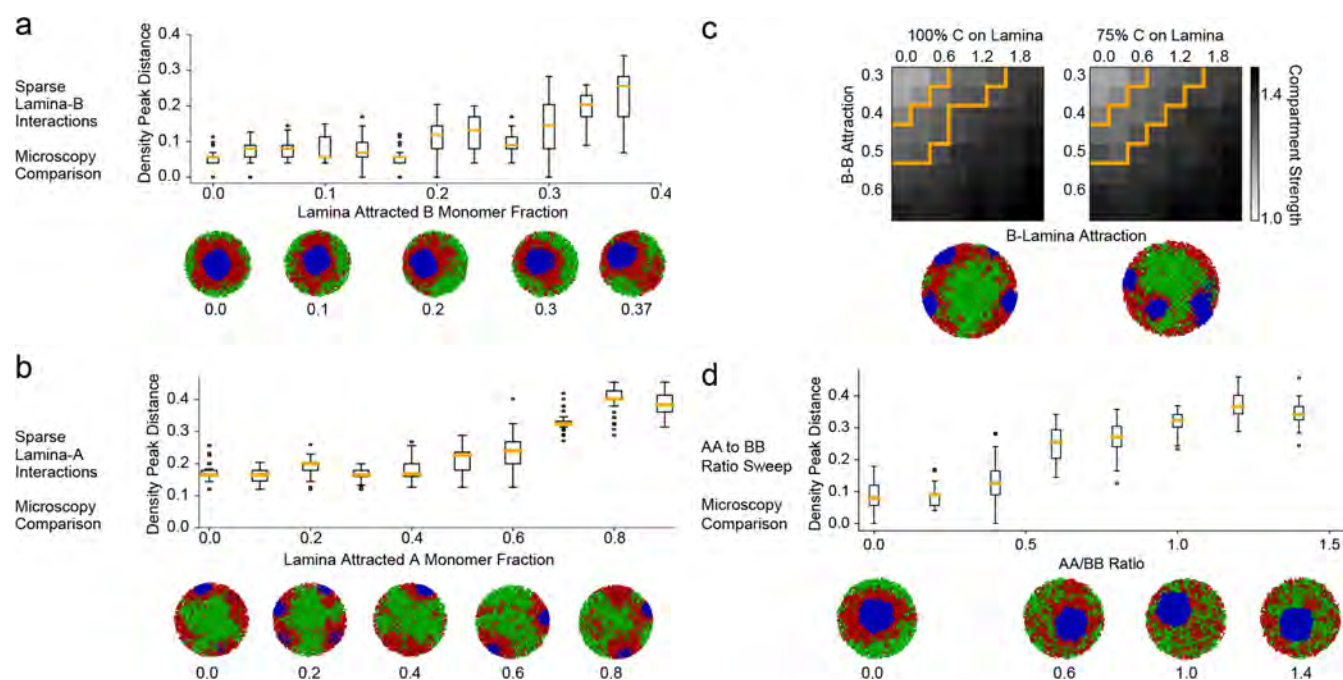
approaching that of biological replicates, which indicates that inversion does not change compartment positions as such. **c–e**, Fractions of loci that remain the same when comparing two different cell types, as well as fractions of loci that switched from B to A and from A to B. The sequence of cell types is taken from the clustering of their compartment profiles. **f**, Compartment strength across multiple mouse cell types (calculated separately for each autosome, $n = 19$ for datasets not considered in main text; $n = 38$ for two replicates of main text datasets. Centre line is the median, the box ranges from the lower to upper quartiles and whiskers extend to $1.5 \times$ the interquartile range.



Extended Data Fig. 6 | Exploring the space of model classes reveals that only a small fraction can reproduce the inverted nuclear geometry.

a, Even the second-best group of models do not display the ring-like structure that is characteristic of the inverted nucleus (the next-best eight models, indicated in pink, after the eight best models described in the main text, which are indicated in gold). Densities are computed from 50 simulated configurations. **b**, In agreement with the Flory–Huggins theory, we find that if the cross-type attraction (for example, A–B) is greater than both of the same-type attractions (A–A and B–B), the two monomer types will not segregate. For models 8, 11 and 15, this is true of both A–B and B–C terms, and as expected, there is mixing between A and B monomers, and B and C monomers in simulations. Similarly, models 9 and 10 have mixed A and C monomers and high A–C attraction; models 12 and 13 have mixed A and B monomers and higher A–B attraction; and model 14 has mixed B and C monomers, with high B–C attraction.

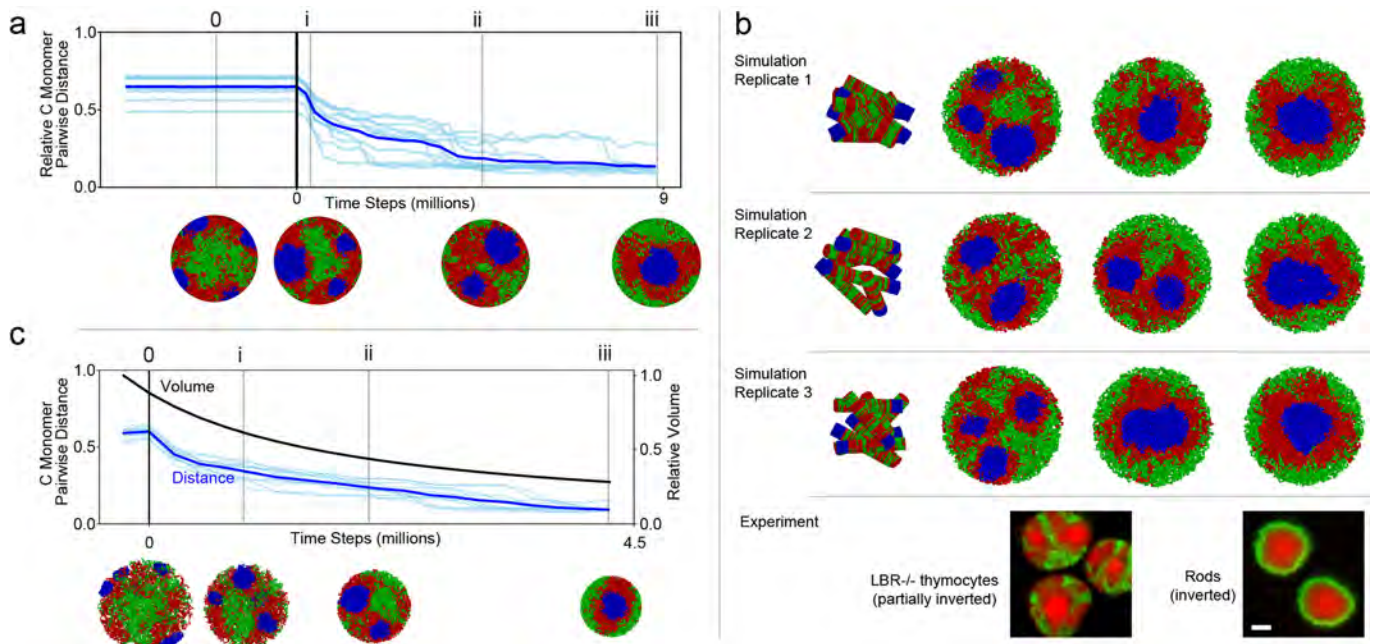
c, Averaging the parameter orders of the second-best model classes reveals that they depart from the best-performing models, in aggregate. **d**, We illustrate particular models with strong euchromatic interactions to show that such models do not compare well with microscopy, even on a quantitative level. In particular, we show the four worst-performing models (pink dots, models 716–719), all of which are characterized by strong euchromatic interactions (**b**). We also show the best-performing model with A–A as its strongest interaction (gold dot, model 250) and the best-performing model with A–A as its second strongest interaction (gold dot, model 61). Neither of these models compare well with experimental microscopy results. Densities are computed from 50 simulated configurations. **e**, All of the poorly performing models discussed in **d** were characterized by strong A–A interactions. **f**, Averaging the worst four models shows that they are characterized by strong A–A interactions.



Extended Data Fig. 7 | The heterochromatin-dominated model is robust to perturbations and outperforms a variety of alternative models.

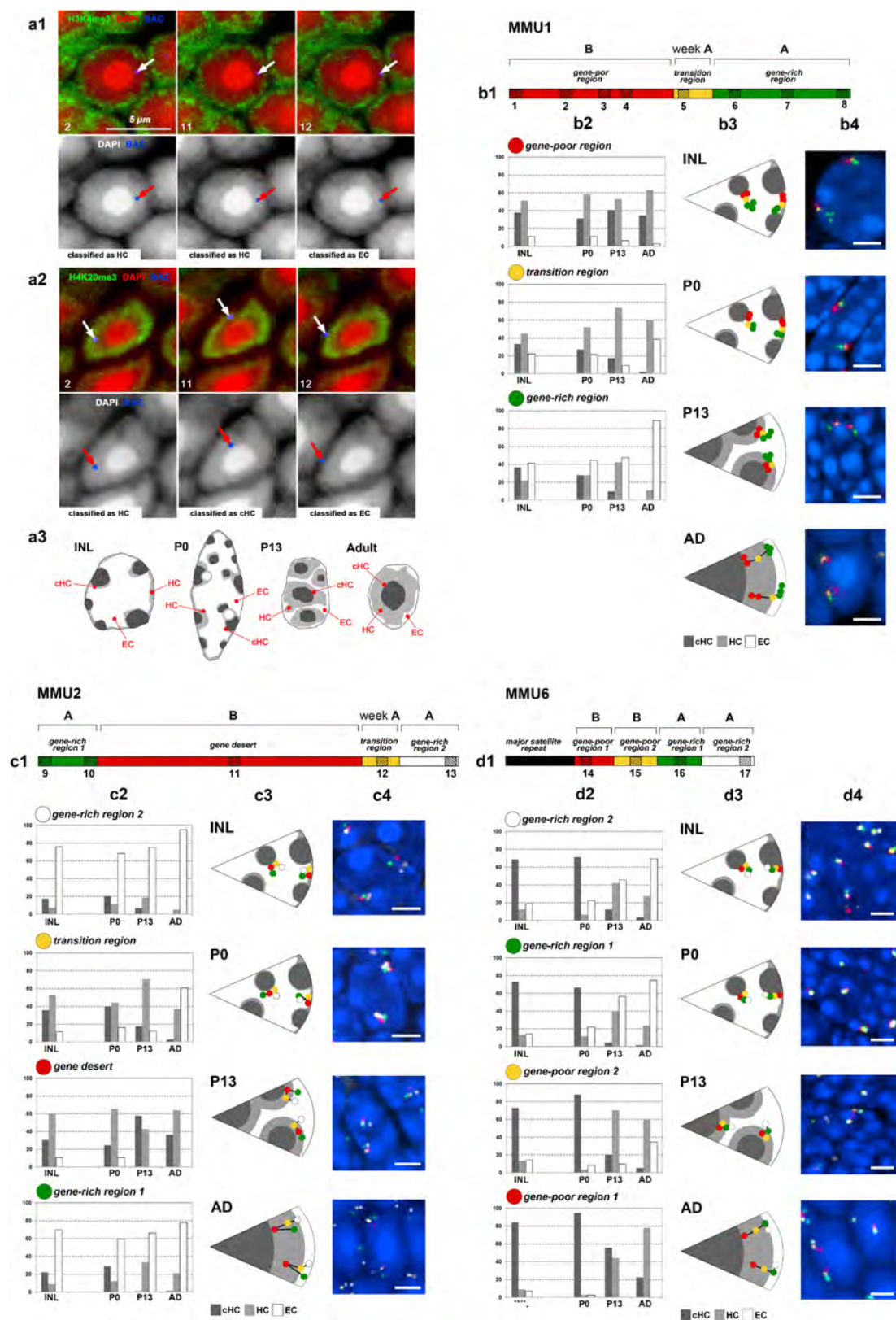
a, Adding in a fraction of B monomers attracted to the lamina, in an analogy to trace amounts of peripheral heterochromatin in rods⁵⁴, does not significantly change agreement with the microscopy results. Representative configurations as this fraction is increased are shown. Boxes indicate density peak distance with whiskers extending to $1.5\times$ the interquartile range. $n = 50$, number of time points sampled across 3 simulation replicates. **b**, Adding in small fractions of A monomers attracted to the lamina (below 20%) does not significantly change the conventional morphology of simulated nuclei. Representative configurations as this fraction is increased are shown. Quantities plotted as in **a**. This simulation reflects a potential phenomenon of association between highly transcribed genes and nuclear pores. Of note, we have not observed this phenomenon in the nuclei of mouse cells, including rod cells, in which all euchromatin is adjacent to the nuclear lamina

(Supplementary Fig. 2). $n = 8$ simulated chromosomes. **c**, Average compartment strength across simulated chromosomes ($n = 8$) as a function of B-B and B-Lam attractions. The zone of parameter space for which simulated Hi-C compartment strength agrees with experimental compartment strength is essentially unchanged for simulations with some interior chromocentres, compared to simulations with no interior chromocentres. Representative configurations of each of these models are displayed. Orange outline indicates regions in parameter space for which the simulated Hi-C has compartmentalization in agreement with experimental Hi-C data (median ± 1 s.d. for wild-type thymocytes). **d**, For B-B = 0.5 and all other parameters as in the main text, increasing the ratio of A-A to B-B results in worse agreement with microscopy. This is particularly visible above A-A/B-B = 0.5. Representative configurations as this fraction is increased are shown. Quantities plotted as in **a**. $n = 8$ simulated chromosomes. Additional models are considered in Supplementary Fig. 6.



Extended Data Fig. 8 | Chromocentres merge during nuclear inversion and pass through a partially inverted morphology. **a**, Distance between chromocentres decreases once interactions with the lamina have been removed, quantitatively showing the fusion of C monomer droplets. To see this, we find the centre of mass of the C monomer blocks on each of the eight chromosomes in our simulation. We then compute the average distance between all possible pairs of the eight centres of mass, and normalize by the maximum possible total separation in the nucleus—that is, the diameter of the nucleus times the number of chromosome pairs. Light-blue lines show individual trajectories, the dark-blue line shows the average over trajectories. Following release from the lamina (vertical black line), this metric decreases, quantitatively confirming what we see visually in the associated configurations (numerals). **b**, Following three

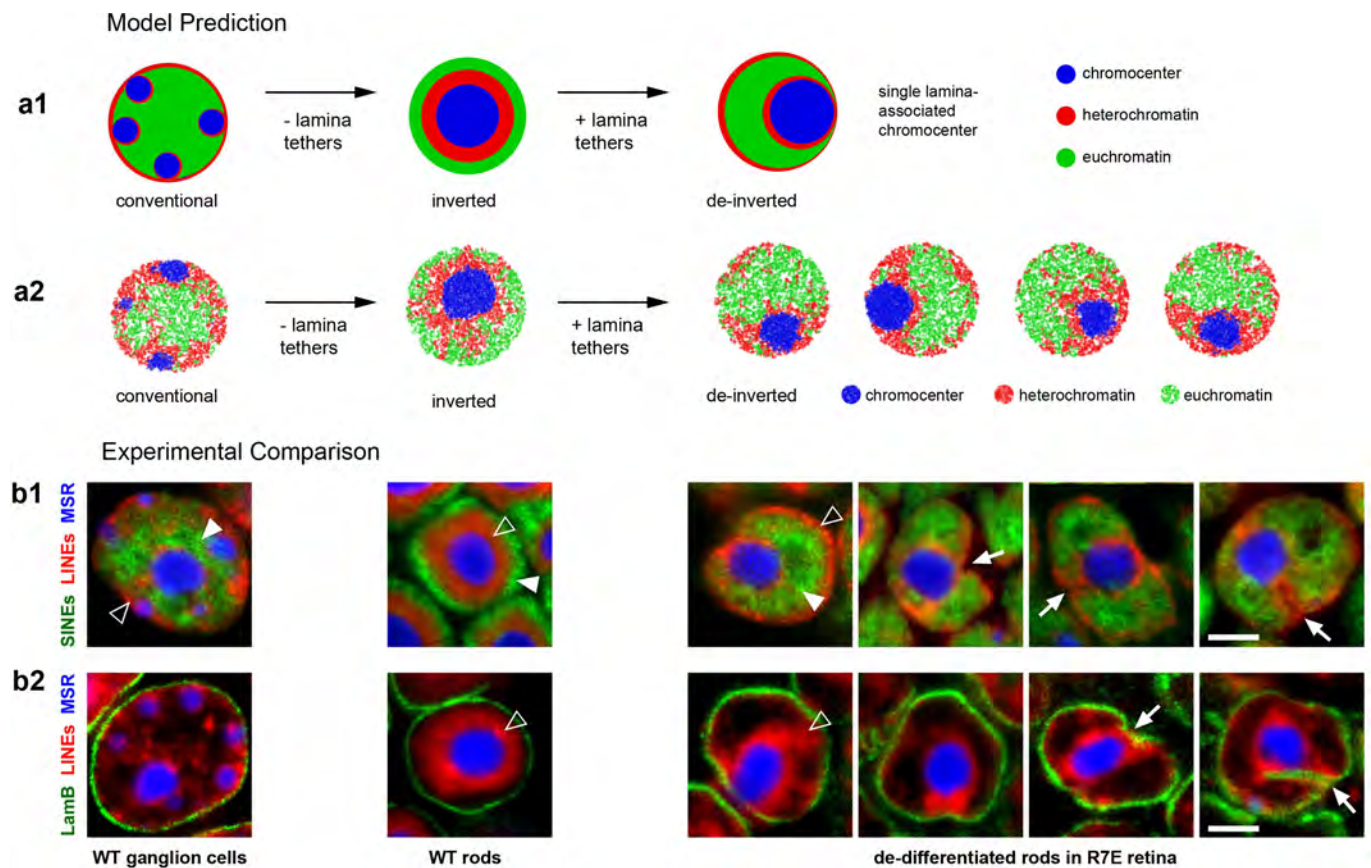
representative simulations starting from an initial condition in which chromosomes are in mitotic-like condensed cylindrical conformations, we find that our inverted nucleus model reaches its equilibrium configuration through a pathway that passes through a state highly reminiscent of the partial inversion seen in *Lbr*^{-/-} thymocytes. As a proxy for detailed mechanistic modelling of the complexities of mitotic exit, we begin from cylinders that are randomly oriented, as opposed to aligned. Scale bar, 2 μm . **c**, Distance between chromocentres decreases once interactions with the lamina have been removed, while the overall volume of the nucleus shrinks at the same time. Quantities plotted as in **a**, with an additional black line for volume decrease relative to initial volume. We see that the qualitative trends in morphology remain the same as in the case of constant volume (Fig. 4a).



Extended Data Fig. 9 | See next page for caption.

Extended Data Fig. 9 | Small chromosome segments faithfully localize to and move together with chromatin of their own compartment during nuclear inversion. The nuclear positions of short chromosome segments of different gene densities belonging to either the A or B compartment were studied using FISH with a cocktail of BAC probes on retinal cryosections at six developmental stages: P0, P6, P13, P21, P28 and adult (AD; 3.5 months). For the analysis of BAC signal distribution, three stages were considered: P0, with conventional nuclei of rod progenitors; P13, with rod nuclei in a transient state of inversion; and adult, with fully inverted rod nuclei. Cells with conventional nuclear organization in the inner nuclear layer (INL) of adult retina were used as a control. Between 100 and 120 alleles per chromosomal region were analysed. **a**, Immuno-FISH experiment showing how FISH signals were classified according to their localization in the three major nuclear zones. EC, euchromatin; HC, heterochromatin; cHC, constitutive heterochromatin. Definitions of these three types of chromatin have been previously published¹. BAC 12 maps to the most peripheral euchromatic shell of the rod nucleus stained with anti-H3K4me3 antibody. This nuclear zone is adjacent to the nuclear periphery and contains the genic part of the mouse genome (see Supplementary Fig. 2). BACs 2 and 11 are located in the heterochromatic zone of the nucleus encircling the chromocentre and stained with anti-H4K20me3 antibody. Thus, classification of BAC signals based on DAPI staining is justified by immunostaining of histone modifications and

enables the signal distribution analysis described in **b–d**. Top, localization of BAC signals (blue, white arrows) and histone modifications (green) in DAPI-counterstained nuclei (red). Numbers in the lower left corners indicate the BAC numbers (for their coordinates, see Methods). Bottom, greyscale images of DAPI and positions of the BAC signals (red arrows) represented by false-coloured mask. **b–d**, Analysis of BAC signal positions after FISH with BAC cocktail probes mapping to selected chromosome regions. Top, schematics of the chromosome regions on MMU1 (**b**), MMU2 (**c**) and MMU6 (**d**). The coloured segments differ in their gene content and assignment to either the A or B compartment. The striped boxes with numbers below indicate the BACs used for FISH. Bottom left, graphs showing the distribution of the segments within rod nuclei at the three developmental stages and adult cells of the inner nuclear layer. The bars represent the proportion of signals in each nuclear zone: adjacent to constitutive heterochromatin (dark grey), within heterochromatin (light grey) and within euchromatin (white). Bottom middle, schematics showing typical segment distribution of the studied regions. Bottom right, representative nuclei after three-colour (**b**) or four-colour (**c**, **d**) FISH. The images are maximum-intensity projections of short (1.4–2- μ m) stacks. False colours assigned to segments correspond to the colour code used in each panel. The experiment was repeated twice. For an example of the localization of a single gene and its movement together with chromatin of the A compartment during nuclear inversion, see Supplementary Fig. 3.



Extended Data Fig. 10 | Coalescence of individual chromocentres into a large central chromocentre is irreversible. **a**, Top, our model predicts that once nuclei invert and all individual chromocentres merge into a single central chromocentre, the reverse process—that is, resplitting into smaller chromocentres—will not take place after the reintroduction of lamina attractions. Although we expect B monomers to redistribute to the nuclear lamina, we do not expect C monomers of a single globule to reorganize into smaller globules. In this sense, our model predicts that the inversion and formation of the central chromocentre is irreversible. Bottom, simulations of de-inversion of inverted nuclei through the introduction of B–Lam and C–Lam attractions with strengths equal to the optimal B–Lam value from Fig. 3c, d. Note that according to our prediction, de-inverted nuclei only partially return to the conventional geometry. Slices with a thickness of 5% of the nuclear diameter are shown. **b**, In agreement with the model prediction, de-inverted nuclei do not return to a typical conventional architecture, as can be seen in de-differentiated rods of R7E mice expressing poly(Q)-expanded ataxin-7 (see Supplementary Fig. 5a, b for a description of the phenotype). FISH with probes for major

satellite repeats (blue), LINE-rich heterochromatin (red) and SINE-rich euchromatin (green) demonstrates that although euchromatin returns to the nuclear interior (filled arrowheads) and heterochromatin repositions to the lamina (empty arrowheads), a single large chromocentre remains and is typically positioned at the nuclear periphery (top, arrows). Notably, in approximately 30% of the nuclei, the large chromocentre does not relocate to the nuclear periphery but the nuclear lamina (green) makes deep narrow invaginations, contacting the chromocentre (bottom, arrows; see also Supplementary Fig. 5c). The remaining bulky chromocentre is surrounded by LINE-rich chromatin (bottom; empty arrowheads) and is often (71% of nuclei) in contact with the nuclear periphery as a result of deformation of nuclear shape (for more examples, see Supplementary Fig. 5c). For comparison, the two left columns show conventional nuclei of ganglion cells and inverted rod nuclei from a wild-type mouse. Images are single optical sections. Scale bars, 2 μ m; scale bars apply to all images. Probes, FISH and microscopy are described in the Methods. Each experiment was repeated three times.

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see [Reporting Life Sciences Research](#). For further information on Nature Research policies, including our [data availability policy](#), see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

► Experimental design

1. Sample size

Describe how sample size was determined.

We used 2 biological replicates for Hi-C analysis of all cell types (rods, non-rod neurons, WT and LBR-null thymocytes).
For each rod cell and non-rod neuron replicate (1 - 10 mln cells), we used retinas from 2 and 3 mice, respectively.
For each thymocyte replicate (40 - 60 mln cells) we used thymus from 1 mouse.

2. Data exclusions

Describe any data exclusions.

No data were excluded from the analysis

3. Replication

Describe whether the experimental findings were reliably reproduced.

All replication attempts were successful

4. Randomization

Describe how samples/organisms/participants were allocated into experimental groups.

We did not perform experiments requiring sample randomization. Differences between data sets were stark enough that visual identification was obvious.

5. Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

Experiments were not blinded. Differences between data sets were stark enough that visual identification was obvious upon viewing data.

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

n/a Confirmed

- ☐ ☒ The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
- ☐ ☒ A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- ☐ ☒ A statement indicating how many times each experiment was replicated
- ☐ ☒ The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)
- ☐ ☒ A description of any assumptions or corrections, such as an adjustment for multiple comparisons
- ☐ ☒ The test results (e.g. P values) given as exact values whenever possible and with confidence intervals noted
- ☐ ☒ A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
- ☐ ☒ Clearly defined error bars

See the web collection on [statistics for biologists](#) for further resources and guidance.

► Software

Policy information about [availability of computer code](#)

7. Software

Describe the software used to analyze the data in this study.

Software used to store and analyze Hi-C data can be accessed at <https://bitbucket.org/mirnylab/hiclib> and <https://bitbucket.org/mirnylab/mirnylib>. Data was also stored using the Cooler software (<https://github.com/mirnylab/cooler>). Simulations were performed with OpenMM v7.2.2 and CUDA v9.1.85. Analysis of data was done with NumPy v1.15.1.

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* [guidance for providing algorithms and software for publication](#) provides further information on this topic.

► Materials and reagents

Policy information about [availability of materials](#)

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.

No unique materials have been used.

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).

=====
Primary antibodies
=====
H4K8ac (72A9, 5.8 µg/µl, 1:200), anti-H4K20me3 (27F10, 5.6 µg/µl, 1:200), H3K9ac (2F3, 3.4 µg/µl, 1:200) - provided by Hiroshi Kimura (described/validated in 10.10 (38/jhg.2013.66 and 10.1007/s10577-013-9375-7)
H3K4me3 (Abcam, ab8580, 1:100)
H3S10ph (Abcam, ab14955, 1:1000)
Ki67 (Abcam, ab15580, 1:200)
Rhodopsin (Abcam, ab3267, clone RET-P1, 1:500)
Lamin A/C (LAZ, serum, undiluted) - provided by Harald Herrmann (described/validated in 10.1007/s00109-007-0275-1)
Lamin B1 (Santa Cruz Biotechnology, SC6217, clone M-20, 1:100)
Nup153 (Abcam, ab24700, QE5, 1:100)
ATAXN7 (1261, 1:100) - provided by Didier Devys (described/validated in 10.1093/hmg/ddh139)
=====
Secondary antibodies
=====
Mouse-anti-Dig antibody conjugated to FITC (Jackson Immuno Research, 200-092-156, 1:100)
Donkey-anti-mouse conjugated to Alexa488 (Invitrogen, A11001, 1:500)
Donkey-anti-mouse conjugated to Alexa555 (Invitrogen, A31570, 1:500)
Donkey-anti-mouse conjugated to Alexa594 (Invitrogen, A21203, 1:500)
Donkey-anti-mouse conjugated to Alexa647 (Invitrogen, A31571, 1:500)
Donkey-anti-rabbit conjugated to Alexa488 (JacksonImmuno Research, 711-547-003, 1:500)
Donkey-anti-rabbit conjugated to Alexa555 (Invitrogen, A31572, 1:500)
Donkey-anti-rabbit conjugated to Alexa649 (JacksonImmuno Research, 711-496-152, 1:500)

10. Eukaryotic cell lines

- State the source of each eukaryotic cell line used.
- Describe the method of cell line authentication used.
- Report whether the cell lines were tested for mycoplasma contamination.
- If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by [ICLAC](#), provide a scientific rationale for their use.

No eukaryotic cell lines were used in this study.

No eukaryotic cell lines were used in this study.

No eukaryotic cell lines were used in this study.

No eukaryotic cell lines were used in this study.

► Animals and human research participants

Policy information about [studies involving animals](#); when reporting animal research, follow the [ARRIVE guidelines](#)

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.

(1) For sampling of rods and WT thymocytes for Hi-C, we used female adult mice (retired breeders) of CD1 strain (<https://www.criver.com/products-services/find-model/cd-1-igs-mouse?region=23>).
(2) For sampling of non-rod neurons for Hi-C, we used female adult mice (2 month old) of C3H strain (<https://www.criver.com/products-services/find-model/c3h-mouse?region=23>).
(3) For sampling of LBR-null thymocytes for Hi-C, we used female adult (3 and 4 month old) LBR GT/GT mice (10.1093/hmg/ddn191).
(4) For the FISH and immunostaining experiments we used retinas of CD1 mice at different stages of development (P0, P3, P6, P13, P21, P28) and adults (14 weeks).
(5) For immuno-FISH in ATAXN7 retinas, retina samples from R7E mice (4, 6, 20, 70 and 93 week old) were provided by Didier Devys (10.1371/journal.pbio.0040067; 10.1016/j.cell.2013.01.009)
Other details are given in the Methods section.

Policy information about [studies involving human research participants](#)

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.

The study did not involve human participants

Late steps in bacterial translation initiation visualized using time-resolved cryo-EM

Sandip Kaledhonkar^{1,5}, Ziao Fu^{2,5}, Kelvin Caban^{3,5}, Wen Li¹, Bo Chen¹, Ming Sun⁴, Ruben L. Gonzalez Jr^{3*} & Joachim Frank^{1,4*}

The initiation of bacterial translation involves the tightly regulated joining of the 50S ribosomal subunit to an initiator transfer RNA (fMet-tRNA^{fMet})-containing 30S ribosomal initiation complex to form a 70S initiation complex, which subsequently matures into a 70S elongation-competent complex. Rapid and accurate formation of the 70S initiation complex is promoted by initiation factors, which must dissociate from the 30S initiation complex before the resulting 70S elongation-competent complex can begin the elongation of translation¹. Although comparisons of the structures of the 30S^{2–5} and 70S^{4,6–8} initiation complexes have revealed that the ribosome, initiation factors and fMet-tRNA^{fMet} can acquire different conformations in these complexes, the timing of conformational changes during formation of the 70S initiation complex, the structures of any intermediates formed during these rearrangements, and the contributions that these dynamics might make to the mechanism and regulation of initiation remain unknown. Moreover, the absence of a structure of the 70S elongation-competent complex formed via an initiation-factor-catalysed reaction has precluded an understanding of the rearrangements to the ribosome, initiation factors and fMet-tRNA^{fMet} that occur during maturation of a 70S initiation complex into a 70S elongation-competent complex. Here, using time-resolved cryogenic electron microscopy⁹, we report the near-atomic-resolution view of how a time-ordered series of conformational changes drive and regulate subunit joining, initiation factor dissociation and fMet-tRNA^{fMet} positioning during formation of the 70S elongation-competent complex. Our results demonstrate the power of time-resolved cryogenic electron microscopy to determine how a time-ordered series of conformational changes contribute to the mechanism and regulation of one of the most fundamental processes in biology.

Initiation of translation is a fundamental step in gene expression that is essential for the overall fitness and viability of cells. In bacteria, the dynamic initiation reaction is kinetically controlled by three initiation factors (IF1; the guanosine triphosphatase IF2; and IF3), which collaborate to ensure accurate selection of fMet-tRNA^{fMet} and its pairing with the mRNA start codon^{10–13}. Canonical initiation begins with assembly of the 30S initiation complex (IC), followed by IF2-catalysed joining of the 50S subunit to the 30S IC to form a 70S IC, and finally maturation of the 70S IC into a 70S elongation-competent complex (EC)^{1,14,15}. Given the essential nature of this process, structural intermediates that form during initiation in bacteria represent promising targets for the development of next-generation antibiotics^{16–18}.

Ensemble rapid kinetic and single-molecule studies have led to the identification and characterization of several intermediate steps during the late stages of initiation. These studies have shown that subunit joining triggers the rapid hydrolysis of GTP by IF2^{1,14,19–21}, dissociation of the initiation factors^{1,12,19}, transition of the ribosomal subunits into their non-rotated inter-subunit orientation^{22,23}, and accommodation of fMet-tRNA^{fMet} into the peptidyl-tRNA binding site of the peptidyl transferase centre (PTC)²⁴. In addition, structures of various 30S^{2–5,18} and 70S ICs^{4,6–8} obtained by cryogenic electron microscopy (cryo-EM)

have revealed intermediate ICs that vary in the conformation of the ribosome, initiation factors and fMet-tRNA^{fMet}. Nonetheless, notable discrepancies in the inter-subunit orientation of the ribosome and position of fMet-tRNA^{fMet} in several of the available structures of the 70S IC have made it difficult to arrive at a consensus structural model for initiation^{4,6–8}. Furthermore, the 70S ICs represented by the available structures were formed using a 70S ribosome and an IF2 bound to a non-hydrolysable GTP analogue (for example, GDPNP) that results in a biochemically trapped 70S IC, rather than by mixing the 50S subunit with a 30S IC that carries a native, GTP-bound IF2 and results in the formation of a 70S IC, which subsequently matures into a 70S EC. Consequently, the available 70S IC structures do not provide information about how the various structural intermediates that have been observed evolve over the course of the initiation reaction. Therefore, these structural studies have been unable to distinguish on-pathway intermediates formed during canonical initiation from spurious, off-pathway intermediates.

To circumvent this problem, and to capture transient, on-pathway intermediates that are created during canonical translation initiation, we have used mixing-spraying time-resolved cryo-EM^{9,25–27}. Previously, we have used this method to study the association of vacant 30S and 50S subunits to form 70S ribosomes²⁵, and to visualize transient structural intermediates formed during the ribosome recycling process²⁶. Having demonstrated that passage through the microfluidic device does not damage the 30S IC (Methods and Extended Data Fig. 1a, b), here we have used mixing-spraying time-resolved cryo-EM to investigate the initiation-factor-catalysed joining of the 30S IC with the 50S subunit to form a transient 70S IC that matures into a 70S EC. Using this approach, we have visualized, in real time and with near-atomic spatial resolution, the conformational rearrangements of the 30S and 70S ICs that promote and control subunit joining, initiation factor dissociation, and fMet-tRNA^{fMet} positioning during 70S EC formation.

Ensemble rapid kinetic studies suggest that transient intermediates formed during initiation are populated on the sub-second timescale^{10,12,14,19–21}. Using published rate constants¹⁹, we developed a kinetic model and analysed how the populations of the expected structural species were predicted to vary as a function of time during subunit joining reactions in which 50S subunits were mixed with 30S ICs (Extended Data Fig. 1c and Supplementary Methods). The analysis predicts that the size of the population of 70S ICs carrying a GTP-bound or GDP-P_i-bound IF2 is maximized at approximately 150 ms, and that joining of the 50S subunit to the 30S IC to form a mature 70S EC is around 65% complete within 600 ms. Using a set of microfluidic chips designed²⁷ to provide reaction times of approximately 20 ms, 80 ms, 200 ms and 600 ms in our mixing-spraying time-resolved cryo-EM apparatus (Extended Data Fig. 2), we therefore mixed 50S subunits with 30S ICs and collected images at each time point. At each time point, two-dimensional (2D) classification of the images yielded 30S subunit-like, 50S subunit-like, and 70S ribosome-like particle classes. Subsequently, the particles from the 20-ms, 80-ms, 200-ms and 600-ms time points were combined into two datasets. The first

¹Department of Biochemistry & Molecular Biophysics, Columbia University, New York, NY, USA. ²Integrated Program in Cellular, Molecular and Biophysical Studies, Columbia University, College of Physicians and Surgeons, New York, NY, USA. ³Department of Chemistry, Columbia University, New York, NY, USA. ⁴Department of Biological Sciences, Columbia University, New York, NY, USA. ⁵These authors contributed equally: Sandip Kaledhonkar, Ziao Fu, Kelvin Caban. *e-mail: rlg2118@columbia.edu; jf2192@cumc.columbia.edu

Table 1 | Populations of the 50S subunit, 70S IC and 70S EC obtained after 3D classification

	50S (%)	70S IC (%)	70S EC (%)
20 ms	58.0 ± 0.7	35 ± 0.8	7.1 ± 0.8
80 ms	36.0 ± 3.4	43.4 ± 1.7	20.6 ± 1.7
200 ms	30.8 ± 5.6	12.7 ± 3.4	56.4 ± 3.4
600 ms	28.4 ± 2.1	8.1 ± 2.0	63.6 ± 2.0

Standard deviations were obtained by repeating the 3D classification procedure three times for each time point.

dataset, containing the 30S subunit-like particles, was subjected to 3D classification. The second dataset, containing 50S subunit-like and 70S ribosome-like particles, was subjected to a combination of 3D and 2D classification to sort out the compositional and conformational heterogeneity (Methods and Extended Data Fig. 3). This classification scheme yielded the structures of five distinct classes: (1) a complex containing the 30S subunit, mRNA and fMet-tRNA^{fMet}, but lacking IF1 and IF2; (2) the 30S IC; (3) the 50S subunit; (4) the 70S IC; and (5) the 70S EC. Execution of an independent, masked classification strategy failed to find any additional rare and/or low-population intermediate conformations of the 70S IC or 70S EC, confirming that our classification scheme did not miss any such states (Methods and Extended Data Fig. 4).

Notably, the sizes of populations of the 50S subunit, 70S IC and 70S EC (Table 1) obtained from our classification strategy qualitatively follow the predicted kinetics (Fig. 1a and Extended Data Fig. 1c), with the population of the 50S subunit decreasing as the population of the 70S EC increases from around 20 ms to 600 ms (Extended Data Fig. 1c). Moreover, the particle populations reported in Fig. 1a and Extended Data Fig. 1c and the structures of the corresponding particle classes are robust to the inclusion of up to 20% ‘noise particles’ falsely picked from the background (Extended Data Fig. 5, Supplementary Methods, Supplementary Tables 1 and 2).

Among the five particle classes that we obtained, we selected the 30S IC, 70S IC and 70S EC for further structural analysis (Fig. 1b, c, 2a). The 70S IC structure reported here is obtained by mixing the 50S subunit with a 30S IC carrying a native, GTP-bound IF2, and the 70S EC structure is obtained directly from a 70S IC formed by an initiation-factor-catalysed initiation reaction. The resolutions of the 30S IC, 70S IC and 70S EC were estimated to be 4.2 Å, 4.0 Å and 3.9 Å, respectively, according to a resolution-estimating protocol that avoids overfitting and uses the Fourier shell correlation (FSC) with the 0.143 criterion²⁸ (Extended Data Fig. 6). Molecular dynamics flexible fitting (MDFF)²⁹ was then used to generate structural models of the 30S IC and 70S IC, and rigid-body fitting of previously published structures of the 30S and 50S subunits (Protein Data Bank (PDB) codes 2AVY and 2AW4, respectively) was used to generate a structural model of the 70S EC (Methods).

Analysis of the 70S ICs that are formed within the first 20–80 ms after mixing 50S subunits with 30S ICs shows that all inter-subunit bridges are formed. Moreover, we find that IF1 has also dissociated from these 70S ICs (compare Fig. 2a and b). This observation is important because IF1 occupies a binding site between the cleft of 16S rRNA helix (h) 44, h18, and ribosomal protein uS12 on the 30S subunit that enables turn 1 of IF1, consisting of residues 18–21, to establish contacts with the minor groove of h44. Consequently, dissociation of IF1 relieves a strong steric clash that would otherwise exist between turn 1 of IF1 and 23S rRNA helix (H) 69 of the 50S subunit (Fig. 2c, d). Because inter-subunit bridge B2a is formed by an interaction between h44 and H69, dissociation of IF1 early during subunit joining enables this crucially important inter-subunit bridge to be established rapidly during initial 70S IC formation.

By the time 80 ms has elapsed, the population of the 70S IC has reached its maximum and, by 200 ms, IF2 has dissociated from a notable fraction of this 70S IC population, resulting in the formation of mature 70S ECs, a process that continues through the 600-ms time point and beyond. Notably, the 70S IC that is captured in this study

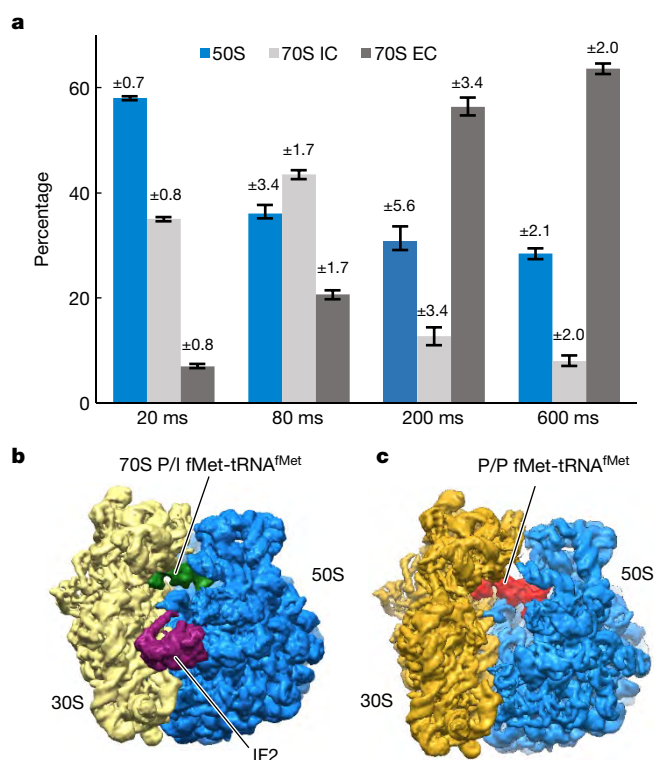


Fig. 1 | Structural and time-resolved population analyses of the 50S subunit, 70S IC and 70S EC. **a**, The populations of the 50S subunit, 70S IC and 70S EC at the 20 ms, 80 ms, 200 ms and 600 ms time points as obtained by 3D classification of the imaged particles. Error bars represent standard deviations obtained by repeating the 3D classification procedure three times for each time point. **b**, **c**, The cryo-EM reconstruction (that is, cryo-EM-derived Coulomb potential maps³⁰) of the 70S IC (**b**) and 70S EC (**c**).

by mixing 50S subunits with 30S ICs carrying native, GTP-bound IF2 is in a semi-rotated inter-subunit orientation that is very similar to the orientation observed in the 70S IC reported previously⁶ and the orientation of the major population of the 70S IC reported in another study⁸ (that is, 70S-IC II). 70S IC-bound IF2 establishes three sets of interactions with the ribosome. Specifically, helix 8 of IF2 interacts with the inter-subunit surface of uS12 in the 30S subunit; domain IV (dIV) of IF2 interacts with H69, H71, H89 and the loop-containing residues 77–85 of ribosomal protein uL16 of the 50S subunit; and domain I (dI) of IF2 interacts with the sarcin-ricin loop (H95) of the 50S subunit (Extended Data Fig. 7). Because the semi-rotated inter-subunit orientation of the 70S ribosome uniquely facilitates the simultaneous formation of these three sets of IF2–ribosome interactions, IF2 selectively stabilizes the 70S IC in this orientation.

Comparative analysis of the 30S and 70S ICs reveals subunit-joining-dependent conformational changes of IF2 that facilitate the formation of the 70S IC. Relative to its position on the 30S IC, dIV of IF2 is stabilized in a position that is approximately 10 Å closer to the 30S subunit—a structural transition that eliminates a potential steric clash with H89 (Fig. 2e, f). Furthermore, the relatively early dissociation of IF1 during 70S IC formation increases the conformational freedom of helix 8 of IF2, allowing it to acquire a position that is closer to the 50S subunit (Fig. 2e). On the basis of previous ensemble rapid kinetic studies^{14,19–21} that demonstrate that the rate of GTP hydrolysis is relatively fast and nearly indistinguishable from the rate of initial subunit association, and that the rate of P_i release is slower than the rate of IF2 dissociation, we propose that we have uniquely captured the native, GDP-P_i-form of IF2 on the 70S IC. This proposal is supported by a structural analysis that demonstrates that GDP-P_i more precisely models the Coulomb potential map³⁰ of the guanosine nucleotide bound to IF2 in the 70S IC than GTP does (Extended Data Fig. 8). The similarity between the conformation of the GDP-P_i-form of IF2 captured here and

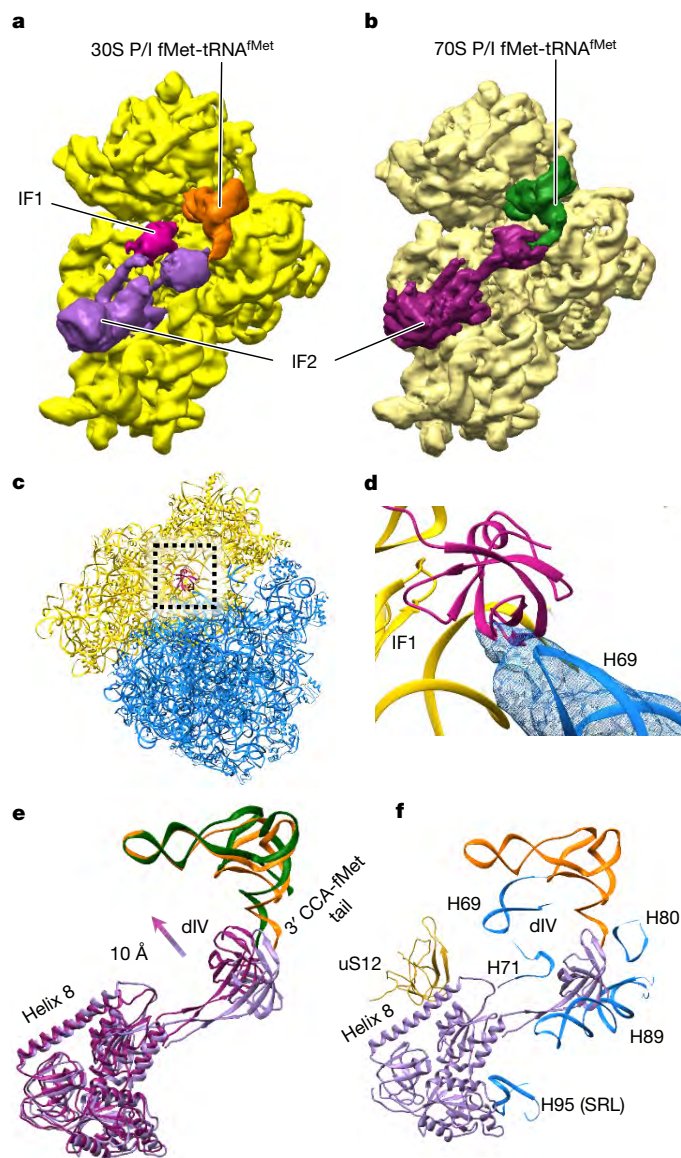


Fig. 2 | Ribosome, initiation factor and fMet-tRNA^{fMet} dynamics during 70S IC formation. **a, b**, Cryo-EM reconstructions viewed from the inter-subunit faces of the 30S IC (with the 30S subunit shown in yellow) (**a**) and the 30S subunit (pale yellow), IF2 and fMet-tRNA^{fMet} from the 70S IC (**b**). **c**, Superposition of the 30S subunits of the 30S IC and the 70S IC and analysis of the conformations of the 30S subunit and IF1 (shown in magenta) from the 30S IC and the 50S subunit from the 70S IC. The analysis reveals that rapid dissociation of IF1 after joining of the 50S subunit to the 30S IC relieves a potential steric clash between IF1 and the 50S subunit that would take place during 70S IC formation. **d**, A magnified view of the superposition shown in **c** highlights the potential steric clash between turn 1 of IF1 and H69 of the 50S subunit. **e**, Superposition of the 30S subunits from the 30S IC and the 70S IC, and comparative analysis of the conformations of IF2 and fMet-tRNA^{fMet} from the 30S IC (light purple and orange, respectively), and IF2 and fMet-tRNA^{fMet} from the 70S IC (dark purple and green, respectively). The analysis reveals that dIV of IF2 moves towards the inter-subunit face of the 30S subunit by approximately 10 Å and, as it rearranges from its 30S P/I to its 70S P/I configuration, the central domain and 3' CCA-fMet tail of fMet-tRNA^{fMet} move slightly towards the tRNA exit (E) site of the 30S subunit after 50S subunit joining to the 30S IC and formation of the 70S IC. **f**, Superposition of the 30S subunits from the 30S IC and 70S IC, and analysis of the conformations of IF2 from the 30S IC and uS12 (shown in pale yellow) of the 30S subunit of the 70S IC and H69, H71, H80, H89 and H95 (the sarcin-ricin loop (SRL); blue) of the 50S subunit of the 70S IC that interact with IF2. The analysis reveals that the IF2 rearrangements shown in **e** relieve a potential steric clash between dIV of IF2 and H89 that would take place during 70S IC formation.

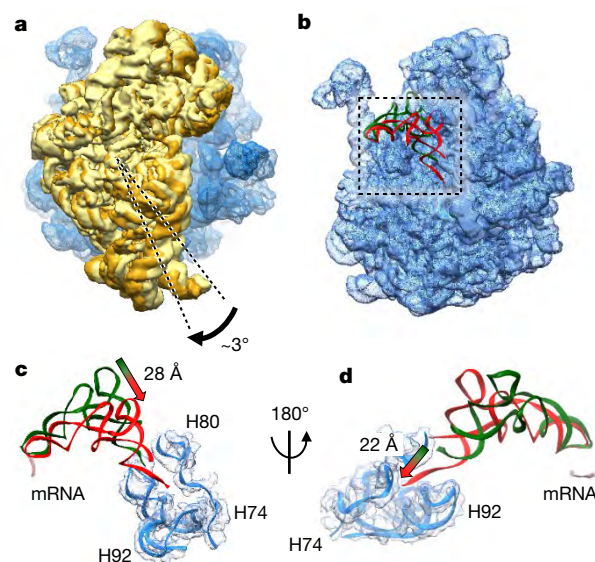


Fig. 3 | Ribosome and fMet-tRNA^{fMet} dynamics during maturation of the 70S IC into a 70S EC. **a**, Superposition of the 50S subunits from the 70S IC and 70S EC and comparative analysis of the conformations of the 30S subunit from the 70S IC (pale yellow) and the 30S subunit from the 70S EC (golden yellow). The analysis reveals that the ribosomal subunits from the 70S IC that is initially formed after 50S subunit joining to the 30S IC transiently acquire a semi-rotated inter-subunit orientation, and subsequently undergo an approximately 3° clockwise rotation, when viewed from the solution side of the 30S subunit, into the non-rotated inter-subunit conformation after maturation of the 70S IC into a 70S EC. **b**, Superposition of the 50S subunits from the 70S IC and 70S EC, and comparative analysis of the conformations of fMet-tRNA^{fMet} in the 70S P/I configuration from the 70S IC (green) and fMet-tRNA^{fMet} in the P/P configuration from the 70S EC (red). The start codon of the mRNA is shown in pale pink. The analysis reveals the conformational rearrangements of fMet-tRNA^{fMet} that take place as the 70S IC matures into a 70S EC. **c**, A magnified view of the superposition shown in **b** reveals that the central domain of fMet-tRNA^{fMet} moves by around 28 Å towards the P site. **d**, A 180° rotation of the superposition shown in **c** highlights the untangling of the 3' CCA fMet tail of the fMet-tRNA^{fMet} and its 22 Å movement into the PTC.

the non-hydrolysable GTP-analogue-form of IF2 reported in all of the other 70S IC structures that have been published^{4,6–8} suggests that when IF2 hydrolyses GTP, it does not immediately undergo a conformational change. This indicates that the transition from the 70S IC to the 70S EC is largely regulated by the release of P_i from IF2 and/or the subsequently rapid release of the GDP-form of IF2 from the 70S IC.

As the 70S IC matures into a 70S EC, dissociation of IF2 disrupts the IF2–ribosome interactions that stabilize the semi-rotated inter-subunit orientation of the 70S IC. Disruption of these IF2–ribosome interactions therefore triggers the reverse rotation of the 30S subunit by approximately 3° (Fig. 3a), which allows the 70S ribosome within the 70S EC to occupy the non-rotated inter-subunit orientation. Dissociation of IF2 also disrupts the contact between dIV of IF2 and fMet-tRNA^{fMet}—an event that, simultaneously with the reverse rotation of the 30S subunit (at least at our time resolution), enables the central domain and 3' CCA-fMet tail of fMet-tRNA^{fMet} to move by around 28 Å and 22 Å, respectively, from the 70S peptidyl/initiation (P/I) configuration of fMet-tRNA^{fMet} that is observed in the 70S IC to the peptidyl/peptidyl (P/P) configuration of fMet-tRNA^{fMet} that is observed in the 70S EC (Fig. 3b–d). This rearrangement of fMet-tRNA^{fMet} is accompanied by an ‘untangling’ of the 3' CCA-fMet tail that allows the fMet moiety to acquire its peptidyl-transfer-competent position within the P site of the PTC (Fig. 3c, d). Given the simultaneous nature of these conformational changes, at least at our time resolution, we propose that the transition of the 70S ribosome into its non-rotated inter-subunit orientation is coupled to the rearrangement

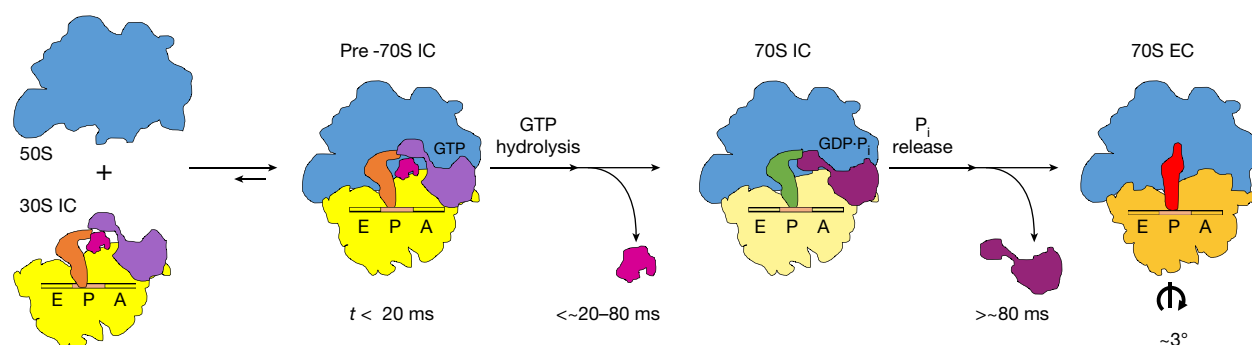


Fig. 4 | Structure-based kinetic model for late steps in bacterial translation initiation. Cartoon depicting the timing of structural and molecular events that occur during the late stages of bacterial translation initiation. Within the first 20 ms after mixing 50S subunits and 30S ICs, 50S subunits (blue) reversibly join to most 30S ICs (yellow) to form transient pre-70S ICs. Conversion of most of these pre-70S ICs into 70S ICs takes place within 20–80 ms after mixing of 50S subunits and 30S ICs and begins with the rapid hydrolysis of GTP on IF2 (light purple). GTP hydrolysis is followed by the dissociation of IF1 (magenta), repositioning of diV of IF2 (dark purple), and formation of IF2–ribosome interactions and inter-subunit bridges that stabilize the ribosome in its

semi-rotated inter-subunit orientation and the fMet-tRNA^{fMet} in its 70S P/P configuration. Within the next several hundred milliseconds, most 70S ICs mature into 70S ECs in a process that begins with the release of P_i from IF2 and dissociation of the GDP-form of IF2 from the 70S IC—events that enable the rotation of the ribosomal subunits into their non-rotated inter-subunit orientation, rearrangement of fMet-tRNA^{fMet} into its P/P configuration, untangling of the 3' CCA-fMet tail of fMet-tRNA^{fMet}, and relocation of the fMet moiety of fMet-tRNA^{fMet} into the PTC in preparation for formation of the first peptide bond after delivery of the first aminoacyl-tRNA into the ribosomal aminoacyl-tRNA binding (A) site.

of fMet-tRNA^{fMet} into its P/P configuration in the 70S EC, along with the untangling of the 3' CCA-fMet tail and positioning of the fMet moiety into the PTC.

Here, we have shown how mixing-spraying, time-resolved cryo-EM is able to capture physiologically relevant, short-lived, structural intermediates in a biomolecular reaction, and have used this approach to determine the molecular mechanism of bacterial translation. On the basis of our collective observations, we propose a structure-based model for the late steps of bacterial translation initiation (Fig. 4). Notably, we did not observe formation of the minor population of the 70S IC reported previously (that is, 70S-IC I)⁸, suggesting that this conformation of the 70S IC might represent an off-pathway intermediate that is formed only when the 70S IC is trapped when using the GDPNP-form of IF2 and/or prepared using a steady-state approach. By contrast, because the conformation of the 70S IC that we observe here was obtained using the native, GTP-bound form of IF2 under pre-steady-state conditions, we can be certain that it represents an intermediate that is formed on the initiation reaction pathway. Mixing-spraying, time-resolved cryo-EM is a new and powerful structural biology technique that we expect will be used to follow the formation and maturation of reaction intermediates and elucidate the molecular mechanisms of fundamental biomolecular reactions such as DNA replication, transcription, precursor mRNA processing and splicing, and mRNA and protein degradation.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests are available at <https://doi.org/10.1038/s41586-019-1249-5>.

Received: 27 July 2018; Accepted: 8 May 2019;
Published online 20 May 2019.

1. Antoun, A., Pavlov, M. Y., Andersson, K., Tenson, T. & Ehrenberg, M. The roles of initiation factor 2 and guanosine triphosphate in initiation of protein synthesis. *EMBO J.* **22**, 5593–5601 (2003).
2. Hussain, T., Llacer, J. L., Wimberly, B. T., Kieft, J. S. & Ramakrishnan, V. Large-scale movements of IF3 and tRNA during bacterial translation initiation. *Cell* **167**, 133–144.e13 (2016).
3. Julián, P. et al. The Cryo-EM structure of a complete 30S translation initiation complex from *Escherichia coli*. *PLoS Biol.* **9**, e1001095 (2011).
4. Simonetti, A. et al. Involvement of protein IF2 N domain in ribosomal subunit joining revealed from architecture and function of the full-length initiation factor. *Proc. Natl Acad. Sci. USA* **110**, 15656–15661 (2013).
5. Simonetti, A. et al. Structure of the 30S translation initiation complex. *Nature* **455**, 416–420 (2008).

6. Allen, G. S., Zavialov, A., Gursky, R., Ehrenberg, M. & Frank, J. The cryo-EM structure of a translation initiation complex from *Escherichia coli*. *Cell* **121**, 703–712 (2005).
7. Myasnikov, A. G. et al. Conformational transition of initiation factor 2 from the GTP- to GDP-bound state visualized on the ribosome. *Nat. Struct. Mol. Biol.* **12**, 1145–1149 (2005).
8. Sprink, T. et al. Structures of ribosome-bound initiation factor 2 reveal the mechanism of subunit association. *Sci. Adv.* **2**, e1501502 (2016).
9. Frank, J. Time-resolved cryo-electron microscopy: Recent progress. *J. Struct. Biol.* **200**, 303–306 (2017).
10. Antoun, A., Pavlov, M. Y., Lovmar, M. & Ehrenberg, M. How initiation factors maximize the accuracy of tRNA selection in initiation of bacterial protein synthesis. *Mol. Cell* **23**, 183–193 (2006).
11. Caban, K. & Gonzalez, R. L. Jr. The emerging role of rectified thermal fluctuations in initiator aa-tRNA- and start codon selection during translation initiation. *Biochimie* **114**, 30–38 (2015).
12. Milon, P., Konevega, A. L., Gualerzi, C. O. & Rodnina, M. V. Kinetic checkpoint at a late step in translation initiation. *Mol. Cell* **30**, 712–720 (2008).
13. Milon, P. & Rodnina, M. V. Kinetic control of translation initiation in bacteria. *Crit. Rev. Biochem. Mol. Biol.* **47**, 334–348 (2012).
14. Grigoriadou, C., Marzi, S., Kirillov, S., Gualerzi, C. O. & Cooperman, B. S. A quantitative kinetic scheme for 70 S translation initiation complex formation. *J. Mol. Biol.* **373**, 562–572 (2007).
15. MacDougall, D. D. & Gonzalez, R. L. Jr. Translation initiation factor 3 regulates switching between different modes of ribosomal subunit joining. *J. Mol. Biol.* **427**, 1801–1818 (2015).
16. Gualerzi, C. O. & Pon, C. L. Initiation of mRNA translation in bacteria: structural and dynamic aspects. *Cell. Mol. Life Sci.* **72**, 4341–4367 (2015).
17. Wilson, D. N. Ribosome-targeting antibiotics and mechanisms of bacterial resistance. *Nat. Rev. Microbiol.* **12**, 35–48 (2014).
18. López-Alonso, J. P. et al. Structure of a 30S pre-initiation complex stalled by GE81112 reveals structural parallels in bacterial and eukaryotic protein synthesis initiation pathways. *Nucleic Acids Res.* **45**, 2179–2187 (2017).
19. Goyal, A., Belardinelli, R., Maracci, C., Milon, P. & Rodnina, M. V. Directional transition from initiation to elongation in bacterial translation. *Nucleic Acids Res.* **43**, 10700–10712 (2015).
20. Huang, C., Mandava, C. S. & Sanyal, S. The ribosomal stalk plays a key role in IF2-mediated association of the ribosomal subunits. *J. Mol. Biol.* **399**, 145–153 (2010).
21. Tomsic, J. et al. Late events of translation initiation in bacteria: a kinetic analysis. *EMBO J.* **19**, 2127–2136 (2000).
22. Ling, C. & Ermolenko, D. N. Initiation factor 2 stabilizes the ribosome in a semirotated conformation. *Proc. Natl Acad. Sci. USA* **112**, 15874–15879 (2015).
23. Marshall, R. A., Aitken, C. E. & Puglisi, J. D. GTP hydrolysis by IF2 guides progression of the ribosome into elongation. *Mol. Cell* **35**, 37–47 (2009).
24. La Teana, A., Pon, C. L. & Gualerzi, C. O. Late events in translation initiation. Adjustment of fMet-tRNA in the ribosomal P-site. *J. Mol. Biol.* **256**, 667–675 (1996).
25. Chen, B. et al. Structural dynamics of ribosome subunit association studied by mixing-spraying time-resolved cryogenic electron microscopy. *Structure* **23**, 1097–1105 (2015).
26. Fu, Z. et al. Key intermediates in ribosome recycling visualized by time-resolved cryoelectron microscopy. *Structure* **24**, 2092–2101 (2016).
27. Lu, Z. et al. Monolithic microfluidic mixing-spraying devices for time-resolved cryo-electron microscopy. *J. Struct. Biol.* **168**, 388–395 (2009).

28. Chen, S. et al. High-resolution noise substitution to measure overfitting and validate resolution in 3D structure determination by single particle electron cryomicroscopy. *Ultramicroscopy* **135**, 24–35 (2013).
29. Trabuco, L. G., Villa, E., Mitra, K., Frank, J. & Schulten, K. Flexible fitting of atomic structures into electron microscopy maps using molecular dynamics. *Structure* **16**, 673–683 (2008).
30. Wang, J., Liu, Z., Frank, J. & Moore, P. B. Identification of ions in experimental electrostatic potential maps. *IUCrJ* **5**, 375–381 (2018).

Acknowledgements This work was supported by funds to J.F. from the National Institutes of Health (R01 GM 55440 and GM 29169) and to R.L.G. from the National Institutes of Health (R01 GM 084288). K.C. was supported by an American Cancer Society Postdoctoral Fellowship (125201).

Reviewer information *Nature* thanks A. Amunts, Simpson Joseph and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions S.K., Z.F., K.C., B.C., M.S., R.L.G. and J.F. designed the research; K.C. prepared all of the biological reagents; S.K. and Z.F. performed

the time-resolved cryo-EM experiments; S.K., Z.F. and W.L. analysed the data; S.K., Z.F., K.C., R.L.G. and J.F. wrote the manuscript; all eight authors approved the final manuscript.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-019-1249-5>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-019-1249-5>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to R.L.G. or J.F.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

METHODS

Preparation, purification and validation of IC components and the 30S IC.

30S and 50S subunits were purified from the MRE600 *Escherichia coli* strain as previously described, with minor modifications³¹. Tight-coupled 70S ribosomes were isolated by ultracentrifugation of crude ribosomes through a 10–40% sucrose density gradient prepared in ribosome storage buffer (10 mM Tris(hydroxymethyl) aminomethane acetate (Tris-acetate) (pH 7.5 at 4°C), 60 mM NH₄Cl, 7.5 mM MgCl₂, 0.5 mM EDTA, 6 mM 2-mercaptoethanol (BME)). To maximize the purity of our tight-coupled 70S ribosomes and minimize contamination by free 50S subunits, a second round of ultracentrifugation through a 10–40% sucrose density gradient prepared in ribosome storage buffer was added to our standard ribosome purification protocol. Highly pure, tight-coupled, 70S ribosomes were buffer-exchanged into ribosome dissociation buffer (10 mM Tris-acetate (pH 7.5 at 4°C), 60 mM NH₄Cl, 1 mM MgCl₂, 0.5 mM EDTA, 6 mM BME) using a centrifugal filtration device (Amicon Ultra, Millipore) with a 100-kDa molecular mass cut-off to promote the dissociation of ribosomes into 30S and 50S subunits. 30S and 50S subunits were isolated from the dissociated tight-coupled 70S ribosomes by ultracentrifugation through a 10–40% sucrose density gradient prepared in ribosome dissociation buffer. To ensure high purity, 30S and 50S subunits isolated from the first gradient were subjected to a second round of ultracentrifugation through a 10–40% sucrose density gradient prepared in ribosome dissociation buffer. Highly purified 30S and 50S subunits were concentrated and buffer-exchanged into ribosome storage buffer using a centrifugal filtration device with a 100-kDa molecular mass cut-off. After determining the concentration of the 30S and 50S subunits, small aliquots were prepared, flash frozen in liquid nitrogen, and stored at –80°C.

The purity of our 30S and 50S subunits was confirmed by negative staining electron microscopy (EM). In brief, one aliquot of the highly purified 30S subunits was diluted to 50 nM with Tris-polymix buffer (50 mM Tris-acetate (pH 7.5 at room temperature), 100 mM KCl, 5 mM ammonium acetate, 0.5 mM calcium acetate, 5 mM magnesium acetate, 0.1 mM EDTA, 6 mM BME, 5 mM putrescine dihydrochloride, and 1 mM spermidine, free base). Subsequently, 3 µl of this 50 nM 30S subunit solution was applied to a carbon-coated EM grid for 30 s. Any excess sample solution was wicked away from the EM grid using filter paper, thereby generating a thin layer of sample solution on the EM grid. Following this, 3 µl of a 2% solution of uranium acetate in water was applied to the EM grid and the EM grid was incubated for 30 s at room temperature. Excess sample solution was wicked away from the EM grid using filter paper, once again generating a thin layer of sample solution on the EM grid. This uranium acetate, negative staining procedure was repeated two more times and, subsequently, the negatively stained 30S subunits were imaged using a 200 kV F20 cryogenic transmission electron microscope (TEM; FEI). Visual inspection of the images that were obtained revealed a highly uniform set of particles exhibiting the characteristically elongated shape of the 30S subunit, thereby demonstrating the purity of the 30S subunits. Analogous procedures were followed to load, negatively stain, and image the highly purified 50S subunits, with visual inspection of the images revealing a highly uniform set of particles exhibiting the characteristic ‘crown view’ of the 50S subunit, thereby demonstrating the purity of the 50S subunits.

IF1 and the γ -isoform of IF2 containing tobacco etch virus (TEV) protease-cleavable, N-terminal, hexa-histidine (6 \times His) tags were overexpressed in BL21(DE3) cells and purified as described previously³². In brief, 6 \times His-tagged initiation factors were purified by nickel nitrilotriacetic acid (Ni²⁺-NTA) affinity chromatography using a batch-binding and elution protocol. After elution of the 6 \times His-tagged initiation factors, the 6 \times His-tags were removed by adding TEV protease to the purified initiation factors and dialysing the mixture overnight (around 12 h) at 4°C against TEV cleavage buffer (20 mM Tris-HCl (pH 7.5 at 4°C), 200 mM NaCl, 0.1% Triton X-100, and 2 mM BME). IF1 was further purified on a HiLoad 16/60 Superdex 75 prep grade gel filtration column (GE Biosciences), and IF2 was further purified on a HiTrap SP HP cation-exchange column (GE Biosciences). The purified initiation factors were concentrated and buffer exchanged into 2 \times translation factor buffer (20 mM Tris-acetate (pH 7.5 at 4°C), 100 mM KCl, 20 mM magnesium acetate, 10 mM BME) using a centrifugal filtration device (Amicon Ultra, Millipore) with either a 3.5-kDa (IF1) or a 10-kDa (IF2) molecular mass cut-off. Concentrated initiation factors were diluted with one volume of 100% glycerol and stored at –20°C. Before using in 30S IC assembly reactions, the initiation factors were buffer exchanged into Tris-polymix buffer using either a Micro Bio-Spin 6 (IF1) or 30 (IF2) gel filtration spin column (Bio-Rad).

tRNA^{fMet} (MP Biomedicals) was aminoacylated and formylated as described previously³². The yield of fMet-tRNA^{fMet}, which was assessed by hydrophobic interaction chromatography (HIC) on a TSKgel Phenyl-5PW column (Tosoh Bioscience) as described previously³², was approximately 90%. The mRNA used in the 30S IC assembly reaction was chemically synthesized (Thermo Fisher) and is a non-biotinylated variant of the bacteriophage T4 gene product (gp) 32 mRNA that we have used extensively in our single-molecule

fluorescence studies of initiation^{15,31,33,34}. The sequence of this mRNA is: 5′-CAAC CUAACUACUACACAAAUUAAAAAGGAAUAGACAU GUUCAAGUCGA AA AAUCUACUGCU-3′.

The 30S IC was assembled by combining 3.6 µM each of IF1, IF2 and fMet-tRNA^{fMet}, 4.8 µM mRNA, 1 mM GTP, and 2.4 µM of 30S subunits in our optimized Tris-polymix buffer³². The final volume of the 30S IC assembly reaction was 100 µl. IF2, which has been previously shown to protect fMet-tRNA^{fMet} from deacylation³⁵ was added to the 30S IC assembly reaction before fMet-tRNA^{fMet}. To ensure that the 30S IC assembly reaction proceeded in a native, unbiased manner, the 30S subunits were added last. Previous ensemble rapid kinetic- and single-molecule studies have shown that IF3 dissociates before³⁶, or shortly after^{12,15,19,34,37,38}, subunit joining. Notably, several of these studies have shown that IF3 regulates the subunit joining reaction by rendering it reversible, establishing a 30S IC + 50S \rightleftharpoons 70S IC dynamic equilibrium in which destabilization of the 70S IC inhibits maturation of the 70S IC into a 70S EC^{13–15,19,37}. Thus, to ensure formation of a stable 70S IC that could productively mature into a 70S EC, IF3 was not included in the 30S IC investigated here. Assembly reactions were incubated at 37°C for 10 min, chilled on ice for 5 min, flash frozen in liquid nitrogen and stored at –80°C.

To assess whether the 30S IC was stable enough to maintain its integrity during mixing-spraying time-resolved cryo-EM, 2.4 µM of 30S IC in Tris-polymix buffer and an equal volume of Tris-polymix buffer lacking 50S subunits were injected into the microfluidic chip designed to give the longest reaction time (600 ms), mixed and sprayed onto an EM grid as the grid was plunge-frozen in liquid ethane. The grid was subsequently stored in liquid nitrogen, all as described above. When ready, the plunge-frozen grid was imaged with the 300 kV Tecnai Polara F30 TEM (FEI), as described above. Subsequently, 2D classification was used to select 30S subunit-like particles. The selected particles were then subjected to 3D classification, which showed that 75% of the 30S subunit-like particles were 30S ICs and 25% of the 30S subunit-like particles were 30S subunits carrying a P-site fMet-tRNA^{fMet} (or, more likely, a mixture of fMet-tRNA^{fMet} and deacylated tRNA^{fMet}) in its ‘30S P/P’ configuration (Extended Data Fig. 1a, b). The results of this experiment demonstrate that most of the 30S IC remains intact during injection and mixing in the microfluidic chip and spraying onto the EM grid.

Preparation of EM grids and mixing-spraying time-resolved cryo-EM. Quantifoil gold R1.2/1.3 grids³⁹ with 300 mesh size were subjected to glow discharge in H₂ and O₂ for 25 s using a Solarus 950 plasma cleaning system (Gatan) set to a power of 25 W. For each of the four time points, 1.2 µM of 50S subunit in Tris-polymix buffer and 2.4 µM of 30S IC in Tris-polymix buffer were injected into the corresponding microfluidic chip at a rate of 3 µl s^{–1} such that they could be mixed and sprayed onto a glow-discharged grid as previously described²⁵. The final concentration of the 50S subunit and the 30S IC after rapid mixing in our microfluidic chip was 0.6 µM and 1.2 µM, respectively. As the mixture was sprayed onto the grid, the grid was plunge-frozen in liquid ethane. The grid was stored in liquid nitrogen until it was ready to be imaged.

Cryo-EM data collection. Plunge-frozen grids were imaged with a 300 kV Tecnai Polara F30 TEM (FEI). The images were recorded within a defocus range of 1–3 µm on a K2 direct detector camera (Gatan) operating in counting mode with an effective magnification of 29,000 \times at 1.66 Å per pixel. Images were composed of 40 frames that were exposed for a total of 12 s, corresponding to a total dose of 35 e[–] Å^{–2}.

Cryo-EM data processing. A flow-chart of the data processing procedure is shown in Extended Data Fig. 3. Micrographs were first screened at each time point, then 468, 605, 445 and 363 micrographs were selected at 20, 80, 200 and 600 ms, respectively. The beam-induced motion of the sample captured by the images was corrected using the MotionCor2 software program⁴⁰. The contrast transfer function (CTF) of each micrograph was estimated using the CTFFIND4 software program⁴¹. Imaged particles were picked using the Autopicker algorithm included in the RELION 2.0 software program⁴². These particles were first extracted using 2 \times binning of the images and subjected to 2D classification to separate 30S subunit-like, 50S subunit-like, and 70S ribosome-like particles from ice-like and/or debris-like particles picked by the Autopicker algorithm. Exclusion of ice-like and debris-like particles resulted in totals of 79,204, 109,775, 59,350 and 66,979 30S subunit-like, 50S subunit-like, or 70S ribosome-like particles at 20, 80, 200 and 600 ms, respectively. All particles classified as 30S subunit-like were pooled together into a set of 170,864 particles, and all particles classified as 50S subunit-like or 70S ribosome-like were pooled together into a set of 144,504 particles. The reason the 50S subunit-like and 70S ribosome-like particles were pooled together is that some of the 70S ribosome-like particles have the appearance of 50S subunit-like particles in particular viewing directions, so separation of these particles must be deferred to the next step.

The set of 170,864 30S subunit-like particles was subjected to a round of 3D classification, from which we obtained two major subclasses. The first subclass encompassed 86,367 30S ICs and the second encompassed 17,686 30S subunits carrying only a P-site fMet-tRNA^{fMet} (or, more likely, a mixture of fMet-tRNA^{fMet}

and deacylated tRNA^{fMet}) in the P/I configuration. The subclass containing the 30S ICs was further refined without binning the images. The resolution of the refined 30S IC was estimated to be 4.2 Å using a resolution-estimating protocol that avoids overfitting and uses the FSC with the FSC = 0.143 criterion²⁸.

The set of 144,504 combined 50S subunit-like and 70S ribosome-like particles was subjected to 3D classification, from which we obtained two major subclasses. The first subclass encompassed 50S subunits and the second encompassed 70S ribosome-like particles. We found it necessary to subject this second subclass to a second round of 2D classification because of evidence of residual compositional heterogeneity. In this step, some 50S subunits were still found, and separated from 80,138 remaining 70S ribosome-like particles. The 50S subunits obtained from this second round of 2D classification were combined with the 50S subunits from the first round of 3D classification for a total of 50,918 50S subunits.

The whole process of two rounds of 2D classification and 3D classification was repeated three times to estimate the errors associated with classifying the set of particles into 50S subunit and 70S ribosome-like particle populations (Table 1). At this point in the analysis, each 50S subunit and 70S ribosome-like particle was traced back to the time point from which it originated to determine the 50S subunit and 70S ribosome-like particle populations at each time point.

The 70S ribosome-like subclass with 80,138 particles was then subjected to a round of 3D classification from which we obtained two major subclasses. The first subclass encompassed 34,096 70S ICs and the second encompassed 46,042 70S ECs. Again, this third round of 3D classification was repeated three times in order to estimate the errors associated with classifying the set of particles into 70S IC and 70S EC populations (Table 1). The subclasses containing the 70S ICs and 70S ECs were then further refined without binning the images. The Fourier amplitudes of the refined cryo-EM maps were sharpened using the 'postprocess' command in RELION. The resolutions of the 70S IC and 70S EC maps were estimated to be 4.0 Å and 3.9 Å, respectively, using a resolution-estimating protocol that avoids overfitting and uses the FSC with the FSC = 0.143 criterion²⁸.

The percentage of particles in the 50S subunit-, 70S IC-, and 70S EC particle classes at each time point was calculated by summing up the number of particles in each particle class at each time point and subsequently calculating the fraction of particles in each particle class with respect to the total number of particles at each time point. To compare the percentages of particles in the 50S subunit, 70S IC and 70S EC particle classes obtained here with the concentrations of 50S subunits, 70S ICs and 70S ECs predicted by the kinetic modelling, the percentages obtained here were used to calculate the concentration of each particle class, assuming the total concentration of particles in the kinetic modelling was limited to 0.6 µM (that is, the limiting concentration of 50S subunits used in the kinetic modelling) (Extended Data Fig. 1c).

Additional 3D classification to find low-population intermediate conformations from the 70S particle dataset. We used masked 3D classification scheme on a dataset of 80,138 70S ribosome particles to search for rare conformations of 70S IC and 70S EC. In the masked 3D classification scheme, a mask was designed covering densities of IF1, IF2, P/P-configured P-site tRNA and P/E-configured P-site tRNA (Extended Data Fig. 4a–c). The dataset of 80,138 70S particles was subjected to 3D refinement to assign angular positions, and particle alignment was turned off during the masked classification scheme. Three types of class were obtained. The first type of classes encompasses 44% of the particles with density for IF2 and tRNA in the P/I position, and the second type of classes encompasses 48% with density for tRNA in the P/P position. The third type of classes encompasses approximately 8% without any density in the masked region. The 3D refinement of the first, second and third types of classes yielded cryo-EM maps of 70S IC, 70S EC and low-resolution 70S EC, respectively (Extended Data Fig. 4d).

Modelling of the 30S IC, 70S IC and 70S EC structures. We obtained near-atomic resolution models of the 30S IC and 70S IC by using the molecular dynamics flexible fitting (MDFF) method²⁹ (Extended Data Fig. 9) and an atomic, cryo-EM-derived model of a 70S IC (PDB code 3JCJ) as the initial starting model.

Similarly, we obtained an initial, near-atomic resolution model of the 70S EC using rigid-body fitting within the UCSF Chimera software program⁴³ and atomic-resolution models of 70S ribosomes in the non-rotated inter-subunit orientation and lacking any tRNA or mRNA ligands (PDB codes 2AVY and 2AW4). This initial, near-atomic-resolution model of the 70S EC was further refined by subjecting it to the 'jiggle fit' algorithm within the COOT software program⁴⁴ to obtain the final atomic coordinates.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

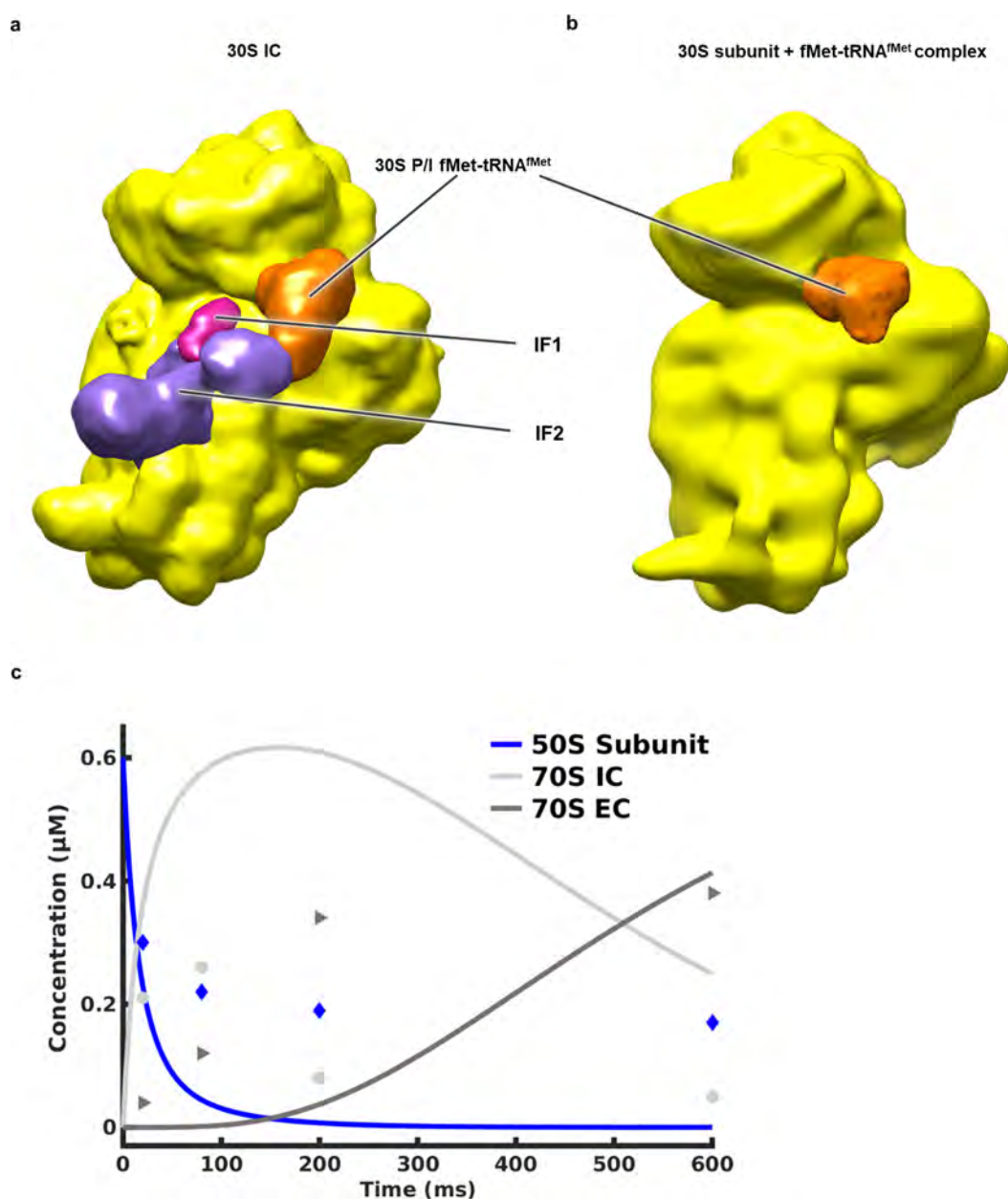
Data availability

The cryo-EM reconstruction maps have been deposited in the Electron Microscopy Data Bank (EMDB) server under the accession codes EMD-0643 (30S IC), EMD-0662 (70S IC) and EMD-0661 (70S EC). The structural models obtained by MDFF have been deposited in the Protein Data Bank (PDB) server under accession codes 6O7K (30S IC) and 6O9K (70S IC). The structural model obtained by rigid-body fitting has been deposited in the PDB server under accession code 6O9J (70S EC).

Code availability

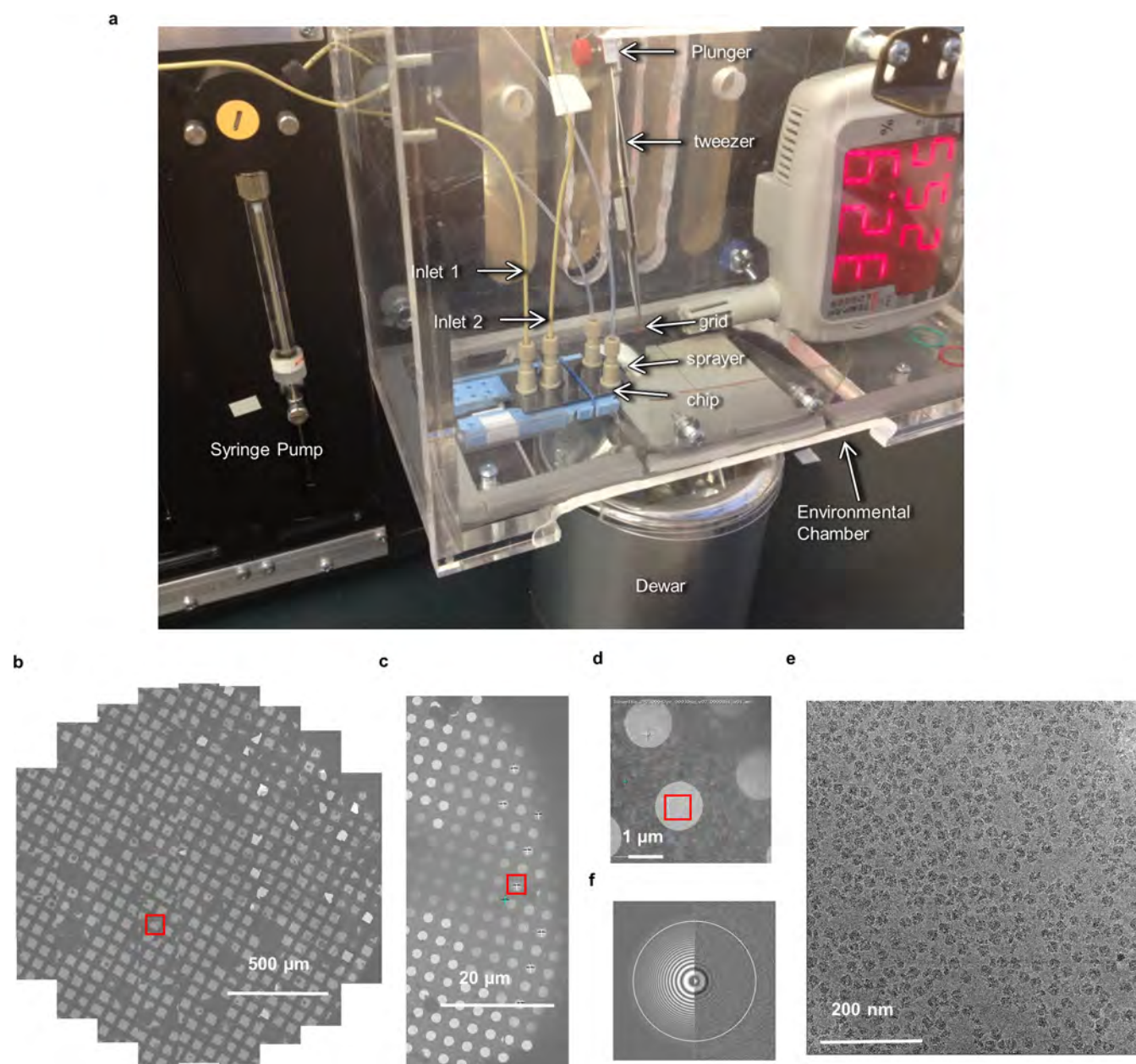
A pseudocode describing the control actions of the software synchronizing time-resolved cryo-EM apparatus is available upon request.

- Caban, K., Pavlov, M., Ehrenberg, M. & Gonzalez, R. L. Jr. A conformational switch in initiation factor 2 controls the fidelity of translation initiation in bacteria. *Nat. Commun.* **8**, 1475 (2017).
- Fei, J. et al. A highly purified, fluorescently labeled in vitro translation system for single-molecule studies of protein synthesis. *Methods Enzymol.* **472**, 221–259 (2010).
- Wang, J., Caban, K. & Gonzalez, R. L. Jr. Ribosomal initiation complex-driven changes in the stability and dynamics of initiation factor 2 regulate the fidelity of translation initiation. *J. Mol. Biol.* **427**, 1819–1834 (2015).
- Elvekrog, M. M. & Gonzalez, R. L. Jr. Conformational selection of translation initiation factor 3 signals proper substrate selection. *Nat. Struct. Mol. Biol.* **20**, 628–633 (2013).
- Guenneugues, M. et al. Mapping the fMet-tRNA^{fMet} binding site of initiation factor IF2. *EMBO J.* **19**, 5233–5240 (2000).
- Antoun, A., Pavlov, M. Y., Lovmar, M. & Ehrenberg, M. How initiation factors tune the rate of initiation of protein synthesis in bacteria. *EMBO J.* **25**, 2539–2550 (2006).
- Grigoriadou, C., Marzi, S., Pan, D., Gualerzi, C. O. & Cooperman, B. S. The translational fidelity function of IF3 during transition from the 30 S initiation complex to the 70 S initiation complex. *J. Mol. Biol.* **373**, 551–561 (2007).
- Fabbretti, A. et al. The real-time path of translation factor IF3 onto and off the ribosome. *Mol. Cell* **25**, 285–296 (2007).
- Russo, C. J. & Passmore, L. A. Electron microscopy: ultrastable gold substrates for electron cryomicroscopy. *Science* **346**, 1377–1380 (2014).
- Zheng, S. Q. et al. MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy. *Nat. Methods* **14**, 331–332 (2017).
- Rohou, A. & Grigorieff, N. CTFFIND4: fast and accurate defocus estimation from electron micrographs. *J. Struct. Biol.* **192**, 216–221 (2015).
- Scheres, S. H. W. RELION: implementation of a Bayesian approach to cryo-EM structure determination. *J. Struct. Biol.* **180**, 519–530 (2012).
- Pettersen, E. F. et al. UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).
- Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D* **60**, 2126–2132 (2004).
- Kaledhonkar, S., Fu, Z., White, H. & Frank, J. in *Protein Complex Assembly: Methods and Protocols* (ed. Marsh, J. A.) 59–71 (Humana, 2018).
- Tan, Y.Z., Baldwin, P.R., Davis, J.H., Williamson, J.R., Potter, C.S., Carragher, B. & Lyumkis, D. Addressing preferred specimen orientation in single-particle cryo-EM through tilting. *Nat. Methods* **14**, 793–796 (2017).
- Raw, A. S., Coleman, D. E., Gilman, A. G. & Sprang, S. R. Structural and biochemical characterization of the GTP_γS-, GDP_γP_γ- and GDP-bound forms of a GTPase-deficient Gly⁴² → Val mutant of G_i1. *Biochemistry* **36**, 15660–15669 (1997).



Extended Data Fig. 1 | Cryo-EM reconstructions. **a, b**, 3D cryo-EM-derived Coulomb potential maps³⁰ of the 30S IC (**a**) and the 30S subunit + fMet-tRNA^{fMet} complex (**b**) obtained from a control experiment in which the 30S IC in Tris-polymix buffer and a solution of Tris-polymix buffer lacking 50S subunits were injected into the microfluidic chip designed to give the longest reaction time (~600 ms), mixed, allowed to react, and sprayed onto an electron microscopy grid that was rapidly plunged into liquid ethane. The sizes of the resulting populations of the 30S IC and the 30S subunit + fMet-tRNA^{fMet} complex were 75% and 25%, respectively, which demonstrates that most of the 30S ICs remain intact during the mixing-spraying process. **c**, Plot of the concentrations of the 50S subunit, 70S IC and 70S EC as a function of time generated by using the initial 50S subunit and 30S IC concentrations analogous to those used in our mixing-spraying microfluidic chip (that is, 0.6 μM and

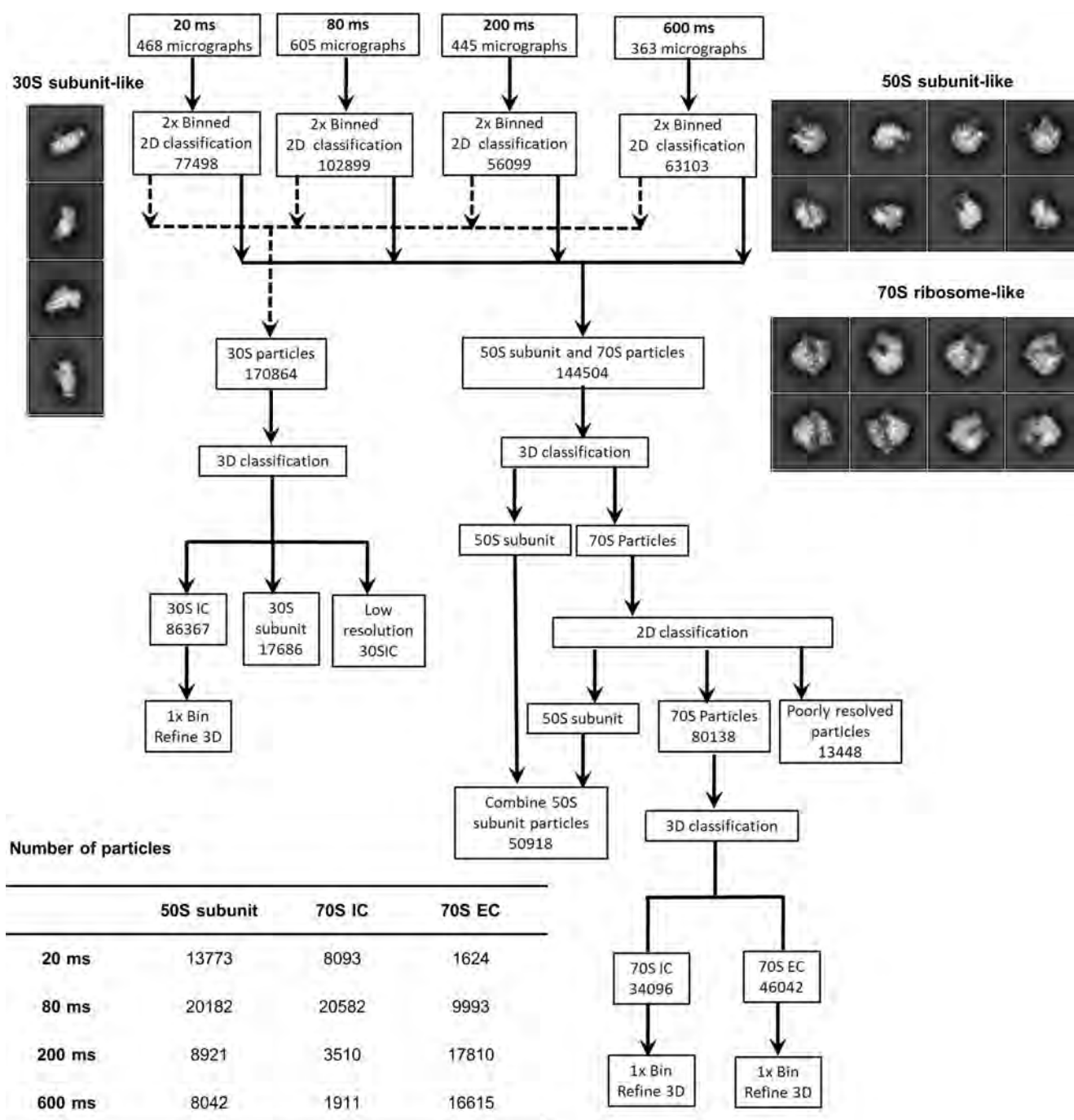
1.2 μM, respectively) and modelling the kinetics of subunit joining using the kinetic scheme and set of rate constants reported previously for a subunit-joining reaction performed in the presence of IF1 and IF2, but in the absence of the IF3¹⁹. A detailed description of the kinetic modelling can be found in the Methods. The plot predicts that the 70S IC population should peak within 50–250 ms after mixing of the 50S subunit and 30S IC, and that these 70S ICs should mature to a notable population of 70S ECs within the next several hundreds of milliseconds. Therefore, to ensure that we would capture formation of the 70S IC and its maturation to the 70S EC, we selected microfluidic chips designed to provide reaction times of approximately 20 ms, 80 ms, 200 ms and 600 ms. The free 50S subunit, 70S IC and 70S EC populations observed in our time-resolved cryo-EM experiments are shown as blue diamonds, light grey circles and dark grey triangles, respectively.



Extended Data Fig. 2 | Time-resolved cryo-EM grid preparation apparatus.

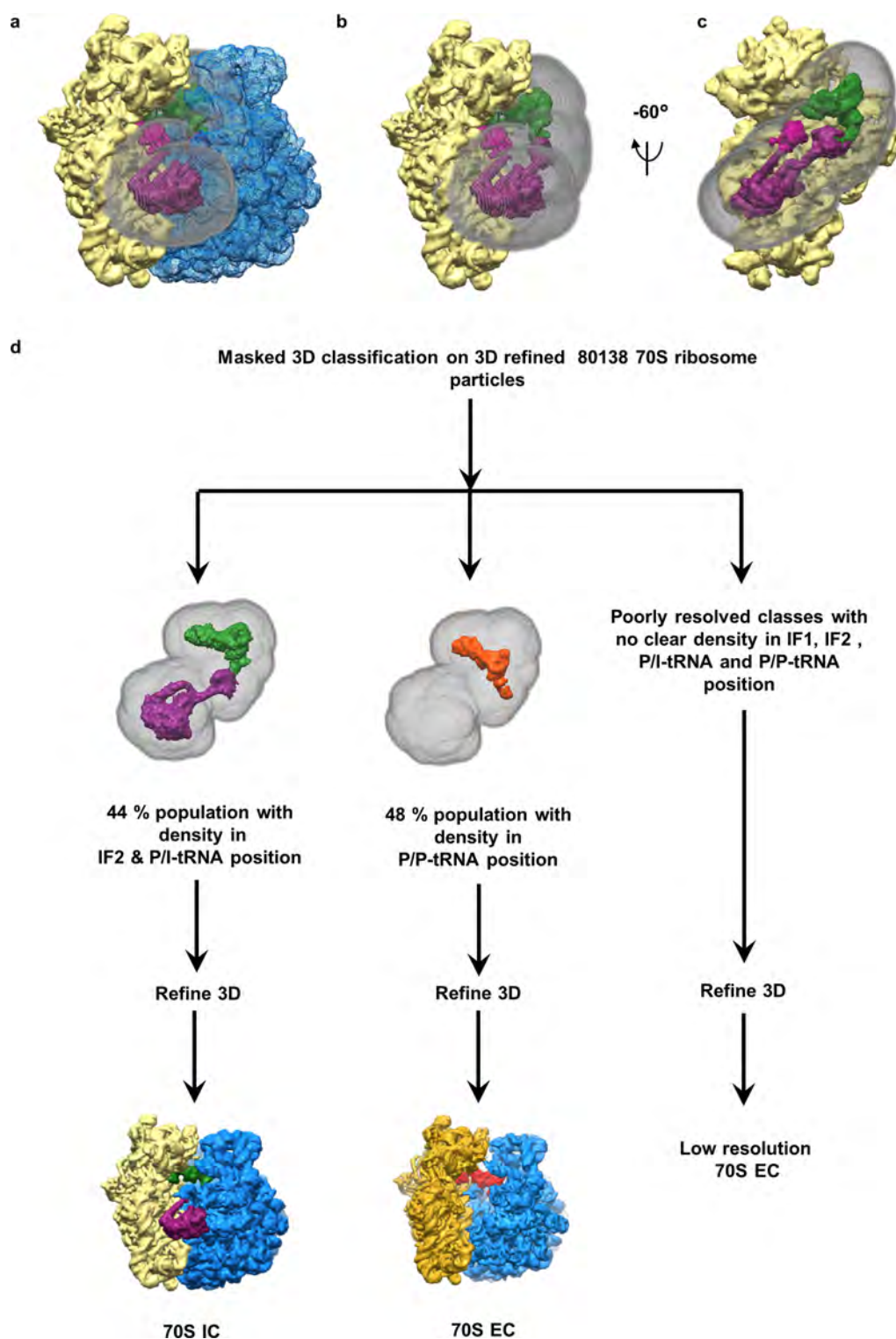
a, A photograph of the mixing-spraying, time-resolved cryo-EM apparatus, labelled to show all major components. The mixing-spraying microfluidic chip is mounted inside an environmentally controlled chamber. A syringe pump, which is controlled by a laboratory-written, Visual Basic and C++ software program called Howard5e⁴⁵, is used to inject the reactants from inlets 1 and 2 into the microfluidic chip. Once in the microfluidic chip, the reactants are mixed and allowed to react for the reaction time specific to the microfluidic chip being used. The electron microscopy grid is held at the end of the plunger by a pair of tweezers. The Howard5e software controls and synchronizes the syringe

pump as well as the plunger that holds the tweezer-mounted electron microscopy grid⁴⁵. Thus, as the sprayers discharge the reaction from the microfluidic chip onto the electron microscopy grid, the plunger is activated to plunge the grid into cryogen⁴⁵. **b–f**, Images of cryo-EM grids prepared by the mixing-spraying, time-resolved cryo-EM apparatus, going from low to high magnification. **b**, Grid-view depicting droplets of different sizes deposited on the grid. **c**, Square-view depicting droplet distribution over the holes. **d**, Hole-view depicting ice distribution over holes. For image acquisition, thin ice regions were selected. **e**, A representative micrograph showing good particle density. **f**, Power spectrum of the acquisition image in **e**.



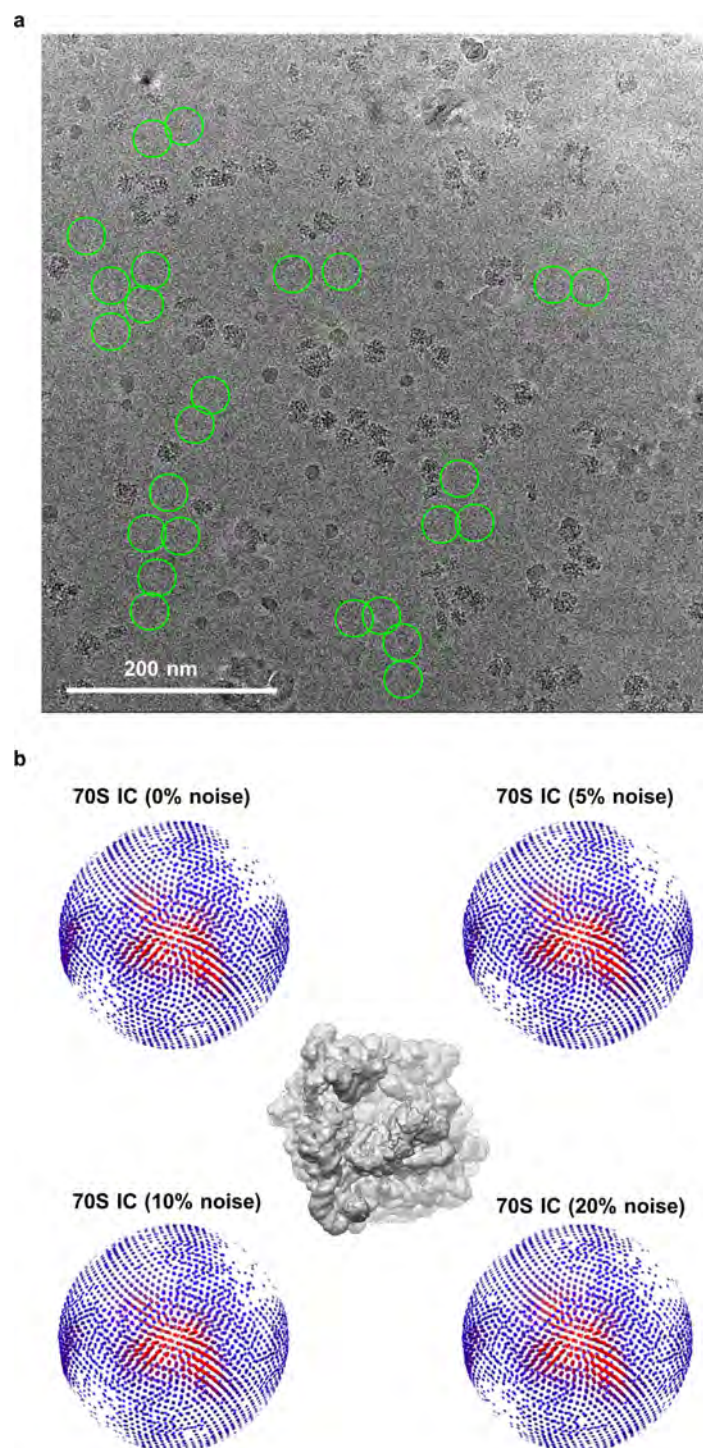
Extended Data Fig. 3 | Flow-chart of the work process for single-particle analysis and 3D refinement. In the first step, particles were auto-picked from the images recorded for the individual time points. Auto-picked particles were then extracted using 2× binning of the images and subjected to 2D classification to discard ice-like and/or debris-like particles and define 30S subunit-like, 50S subunit-like and 70S ribosome-like particle classes. Representative 2D classes of 30S subunit-like, 50S subunit-like and 70S ribosome-like particles are shown on the left and right of the flow chart. A detailed account of the classification scheme is provided in the Methods. In brief, following 2D classification at each time point, two particle datasets were created. The first particle dataset was composed of 170,864 30S subunit-like projections, and the second was composed of 144,504 50S subunit-like and 70S ribosome-like projections. The first particle dataset with 170,864 30S subunit-like

projection classes was subjected to 3D classification, which yielded two major subclasses. The first of these contained 86,367 30S ICs and the second contained 17,686 30S subunits carrying only fMet-tRNA^{fMet} (that is, 30S subunit + fMet-tRNA^{fMet} complexes). The second particle dataset, containing the 144,504 50S subunit-like and 70S ribosome-like particles, was also subjected to a combination of 3D and 2D classification to separate compositional heterogeneity consisting of the 50S subunit ribosome and 70S ribosome. After performing a combination of 3D and 2D classifications, two particle datasets were created, the first containing 50,918 50S subunit particles and the second containing 80,138 70S ribosome-like particles. Further 3D classification was performed on the dataset containing 80,138 70S ribosome-like particles, which yielded 70S IC and 70S EC classes. Particles from 50S subunit, 70S IC and 70S EC were traced back to each time point, as tabulated at the bottom of the flow chart.



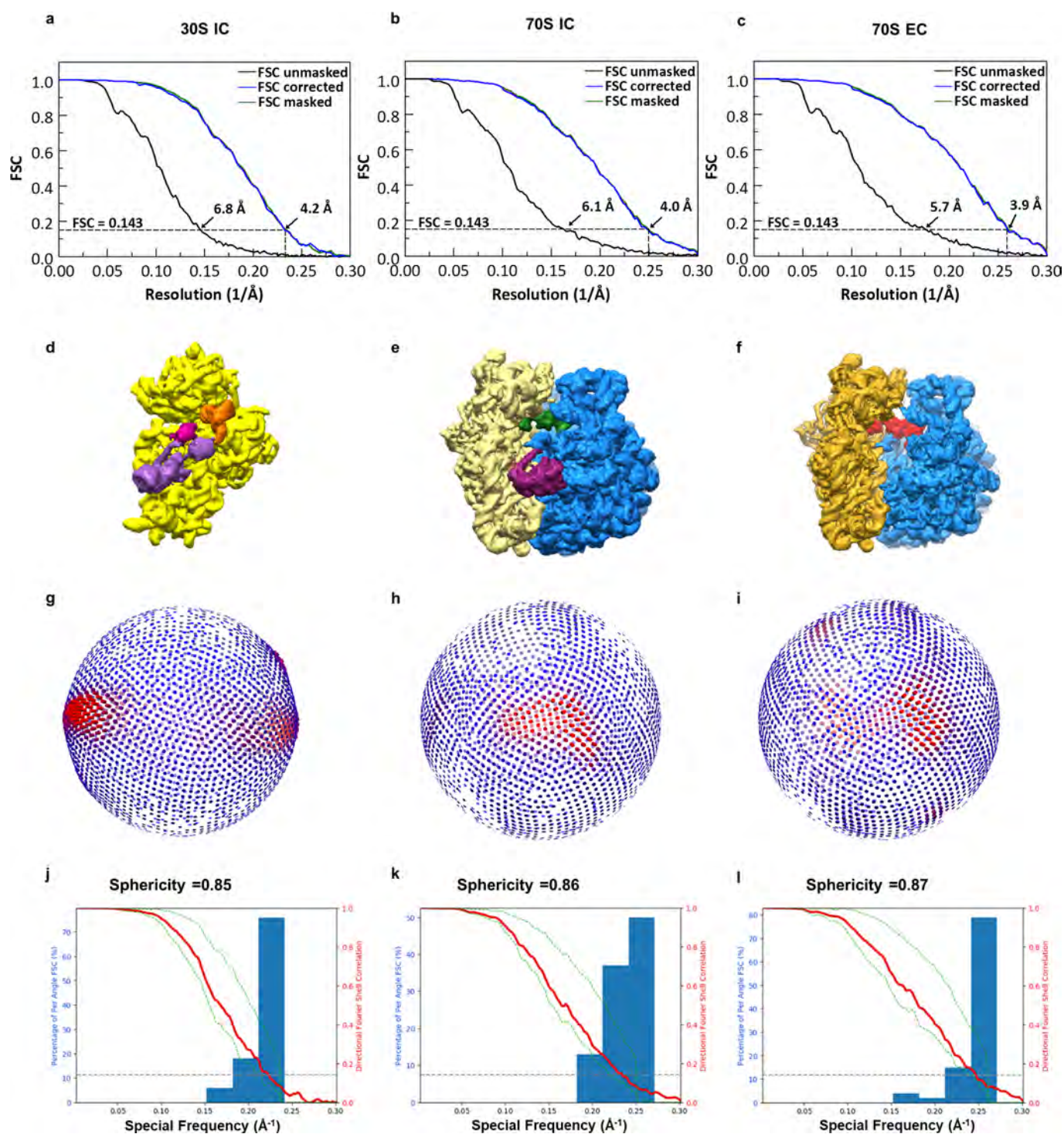
Extended Data Fig. 4 | Masked classification scheme to look for rare conformations of the 70S IC and 70S EC. **a–c**, The mask (grey) covering densities of IF1 (magenta), IF2 (purple), P/P-tRNA (orange) and P/E-tRNA is shown in different views. The views depict the position of the mask with respect to the 30S subunit (pale yellow) and the 50S subunit (blue). **d**, For the masked 3D classification scheme, this mask was applied to the dataset of refined 80,138 70S particles, which yielded mostly three types

of class. The first class encompasses 44% of the particles with density for IF2 (purple) and tRNA in the P/I position (green), the second class encompasses 48% with density for tRNA in the P/P position (orange), and the third class encompasses approximately 8% without any density in the masked region. The 3D refinement of the first, second and third types of class yielded cryo-EM maps of 70S IC, 70S EC and low-resolution 70S EC, respectively.



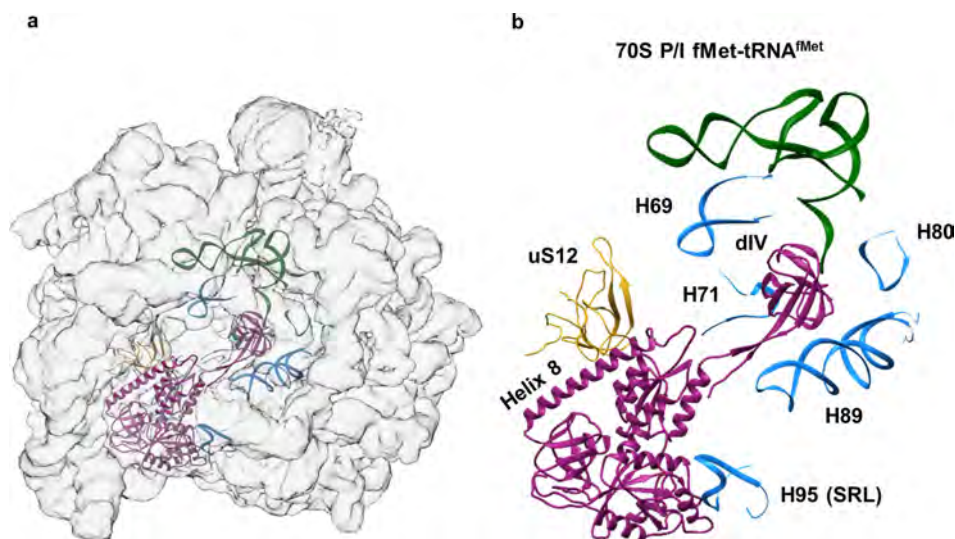
Extended Data Fig. 5 | Selection of noise particles and angular coverage of 70S IC with addition of noisy particles. a, Noise particles selected from the gain-corrected micrograph. The noise particles that did not exhibit any

ribosome particle-like features were selected from the background (green circles). **b,** The angular coverage of the 70S IC with respect to the view depicted in the centre panel, as a function of the level of added noise.



Extended Data Fig. 6 | Fourier shell curves and cryo-EM reconstructions for the 30S IC, 70S IC and 70S EC. a–c, Fourier shell curves (FSC) for the 30S IC (a), 70S IC (b) and 70S EC (c). The resolutions of these structures were estimated using a resolution-estimating protocol that avoids overfitting and uses the FSC and 0.143 criterion²⁸. d–f,

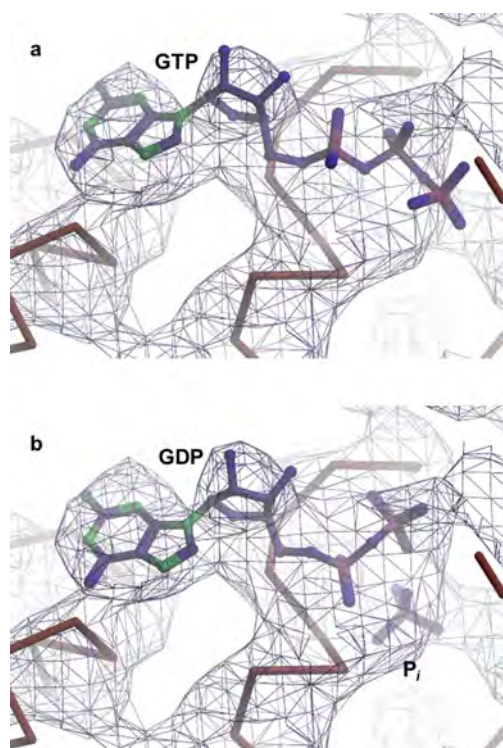
Cryo-EM reconstructions of 30S IC (d), 70S IC (e) and 70S EC (f). g–i, Angular orientation coverage of 30S IC (g), 70S IC (h), and 70S EC (i), presented corresponding to the views depicted in d, e and f, respectively. j–l, Directional FSC plots⁴⁶ of the cryo-EM reconstructions of the 30S IC (j), 70S IC (k) and 70S EC (l).



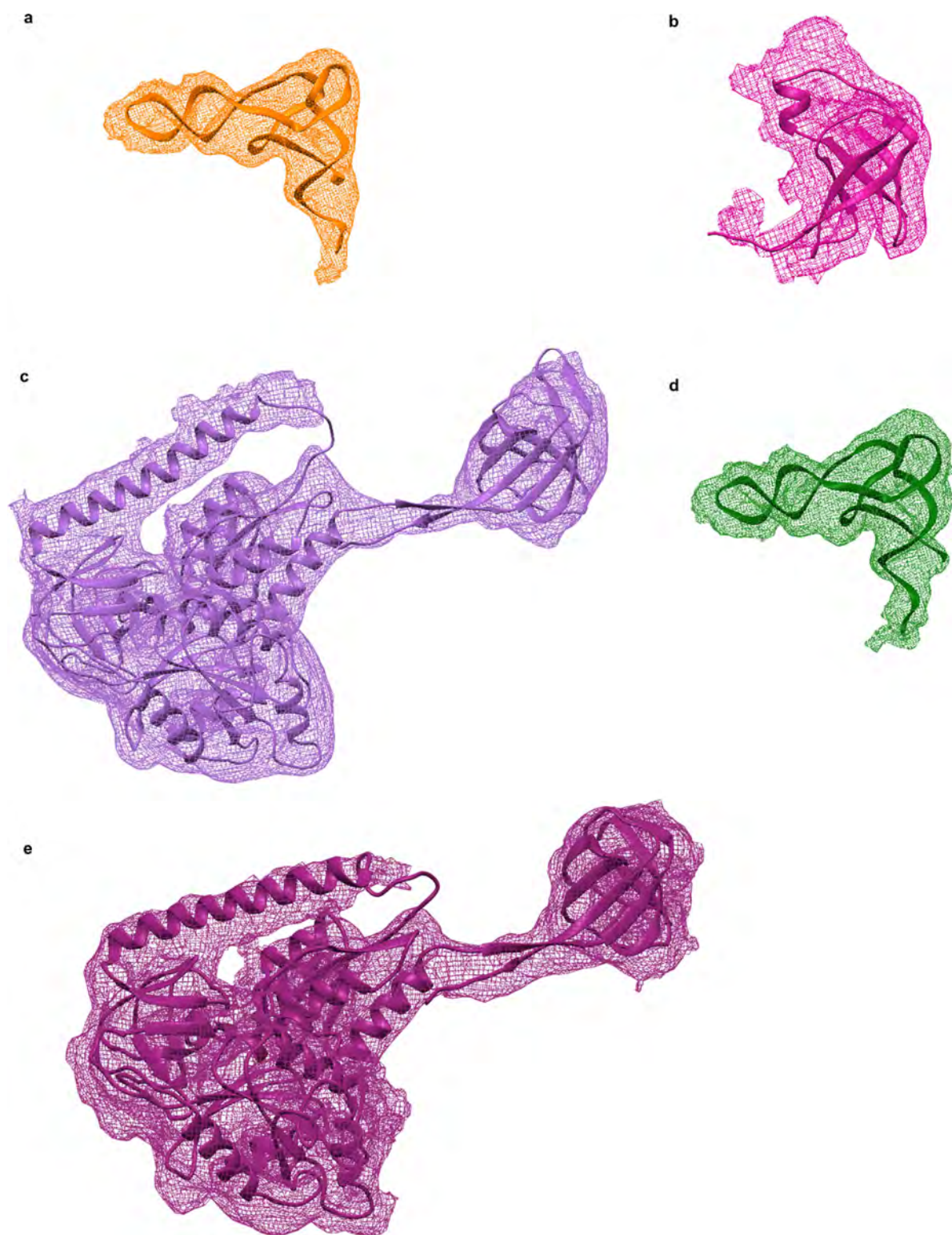
Extended Data Fig. 7 | IF2-70S ribosome interactions in the 70S IC.

a. The positions of IF2 (dark purple), 70S P/I fMet-tRNA^{fMet} (green), uS12 (yellow), and H69, H71, H80, H89 and H95 (SRL; all shown in blue) within the cryo-EM reconstruction of the 70S IC (transparent grey) were

obtained using the MDFF method. **b.** A magnified view of the structure shown in **a**, highlighting the interactions that IF2 makes with the 70S ribosome in the 70S IC.



Extended Data Fig. 8 | Portion of the Coulomb potential map corresponding to the guanosine nucleotide obtained from the 4 Å resolution, cryo-EM reconstruction of the 70S IC. a, b, Rigid-body fitting was used to position either GTP (a) or GDP·P_i (b) into the Coulomb potential map. The Coulomb potential map is shown as a blue mesh. The initial position of P_i relative to GDP for the rigid-body fitting was taken from the structure of the GDP·P_i-form of the G protein G₁₀₁ (PDB code 1AS2)⁴⁷.



Extended Data Fig. 9 | Views of the major components of the 30S IC and 70S IC after structural modelling of the 30S IC and 70S IC using the MDFF method. a, fMet-tRNA^{fMet} (orange) in its 30S P/I configuration in the 30S IC. **b,** IF1 (magenta) in the 30S IC. **c,** IF2 (light purple) in the 30S

IC. **d,** fMet-tRNA^{fMet} (green) in its 70S P/I configuration in the 70S IC. **e,** IF2 (dark purple) in the 70S IC. For each component, the reconstructed Coulomb potential map is represented by the mesh and the structural model is represented by secondary structure cartoons.

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a Confirmed

- ☒ ☐ The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- ☒ ☐ An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- ☒ ☐ The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- ☒ ☐ A description of all covariates tested
- ☒ ☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- ☐ ☒ A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- ☒ ☐ For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
Give P values as exact values whenever suitable.
- ☒ ☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- ☒ ☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- ☒ ☐ Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated
- ☐ ☒ Clearly defined error bars
State explicitly what error bars represent (e.g. SD, SE, CI)

Our web collection on [statistics for biologists](#) may be useful.

Software and code

Policy information about [availability of computer code](#)

Data collection

LEGION

Data analysis

MotionCor2, CTFFIND4, RELION 2.0, Microsoft Excel, UCSF Chimera software v1.12, COOT 0.8.9, VMD, MATLAB 2015b

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The cryo-EM reconstruction maps were deposited in the EMDB server under the accession codes EMD-0643 (30S IC), EMD-0662 (70S IC), and EMD-0661 (70S EC). The structural models obtained by MDFF were deposited in the PDB server under accession codes 6O7K (30S IC) and 6O9K (70S IC). The structural model obtained by rigid-body fitting was deposited in the PDB server under accession code 6O9J (70S EC).

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/authors/policies/ReportingSummary-flat.pdf](https://www.nature.com/authors/policies/ReportingSummary-flat.pdf)

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	The number of particle images were chosen with auto-picker algorithm in RELION 2.0, and amply sufficient for the statistics. All of these strategies has been already published and widely discussed in research conferences.
Data exclusions	The principle of data exclusion, via classification was performed on previously published strategy which is widely accepted
Replication	replication of the data analysis is used routinely as a principle of the resolution test. In our analysis resolution tests yielded standard deviation in the range of 0.7-1.5.
Randomization	this is part of the Relion software approach, and has been discussed at length in published article Scheres, S.H.W. (2012) J. Struct. Biol.
Blinding	Particle selection and classification decisions that were done independently by two team members were in overall agreement

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Unique biological materials
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Human research participants

Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

CORRECTIONS & AMENDMENTS

CORRECTION

<https://doi.org/10.1038/s41586-019-1266-4>

Author Correction: miR-34a blocks osteoporosis and bone metastasis by inhibiting osteoclastogenesis and Tgif2

Jing Y. Krzeszinski, Wei Wei, HoangDinh Huynh, Zixue Jin,
Xunde Wang, Tsung-Cheng Chang, Xian-Jin Xie, Lin He,
Lingegowda S. Mangala, Gabriel Lopez-Berestein,
Anil K. Sood, Joshua T. Mendell & Yihong Wan

Correction to: *Nature* <https://doi.org/10.1038/nature13375>,
published online 25 June 2014.

In this Letter, the citation to ‘Fig. 4e, f’ should be ‘Fig. 3e, f’ following the text: “Consequently, OVX-induced bone loss was attenuated by miR-34a-CH”. This does not affect the conclusions of the paper, and the original Letter has not been corrected online.

CORRECTIONS & AMENDMENTS

CORRECTION

<https://doi.org/10.1038/s41586-019-1221-4>

Author Correction: A dissipatively stabilized Mott insulator of photons

Ruichao Ma, Brendan Saxberg, Clai Owens, Nelson Leung, Yao Lu, Jonathan Simon & David I. Schuster

Correction to: *Nature* <https://www.nature.com/articles/s41586-019-0897-9>, published online 06 February 2019.

Two references^{1,2} were inadvertently omitted from this Article. They have been added as refs ¹² and ¹⁴ at the end of the sentence “Recently, photonic systems have emerged as a platform of interest for the exploration of synthetic quantum matter.” in the original Article. Subsequent references are renumbered accordingly.

1. Hartmann, M. J., Brandão, F. G. S. L. & Plenio, M. B. Strongly interacting polaritons in coupled arrays of cavities. *Nat. Phys.* **2**, 849–855 (2006).
2. Angelakis, D. G., Santos, M. F. & Bose, S. Photon blockade induced Mott transitions and XY spin models in coupled cavity arrays. *Phys. Rev. A* **76**, 031805 (2007).

CORRECTIONS & AMENDMENTS

CORRECTION

<https://doi.org/10.1038/s41586-019-1265-5>

Author Correction: Design of amidobenzimidazole STING receptor agonists with systemic activity

Joshi M. Ramanjulu, G. Scott Pesiridis, Jingsong Yang, Nestor Concha, Robert Singhaus, Shu-Yun Zhang, Jean-Luc Tran, Patrick Moore, Stephanie Lehmann, H. Christian Eberl, Marcel Muelbaier, Jessica L. Schneck, Jim Clemens, Michael Adam, John Mehlmann, Joseph Romano, Angel Morales, James Kang, Lara Leister, Todd L. Graybill, Adam K. Charnley, Guosen Ye, Neysa Nevins, Kamelia Behnia, Amaya I. Wolf, Viera Kasparcova, Kelvin Nurse, Liping Wang, Ana C. Puhl, Yue Li, Michael Klein, Christopher B. Hopson, Jeffrey Guss, Marcus Bantscheff, Giovanna Bergamini, Michael A. Reilly, Yiqian Lian, Kevin J. Duffy, Jerry Adams, Kevin P. Foley, Peter J. Gough, Robert W. Marquis, James Smothers, Axel Hoos & John Bertin

Correction to *Nature* <https://doi.org/10.1038/s41586-018-0705-y>, published online 07 November 2018.

In this Letter, author Ana C. Puhl was inadvertently omitted from the list of authors, with affiliation: Platform Technology & Science, GlaxoSmithKline, Collegeville, PA, USA. At the time this research was initiated, Ana C. Puhl held a post-doctoral position at GlaxoSmithKline through a fellowship awarded from the Brazilian government (the ‘Science without Borders’ programme), and the GlaxoSmithKline ‘Trust in Science’ programme. She performed crystallization trials and was instrumental in the determination of the structure of compounds **1** and **2** described in Figs. 1 and 2. In the ‘Author contributions’ section, the relevant sentence should read: “N.C. performed HDX studies and determined X-ray structures with assistance from L.W. and A.C.P.”. The Letter has been corrected online.

CAREERS

CAREERS TOOLKIT Resources for early-career scientists go.nature.com/2znrii1

WORK-LIFE BALANCE Maintain a healthy footing go.nature.com/2zw9syv

E-NEWSLETTER Sign up at go.nature.com/careersnewsletter

COLLEEN MORGAN



BY EMILY SOHN

Nana Apenem Dagadu has taken her two young children to Senegal, Uganda and Ghana — all locations for her fieldwork as a reproductive- and sexual-health researcher. Her son, aged six, and her daughter, three, have often joined her in the field, partly owing to necessity, and partly because Dagadu enjoys travelling with her kids. She also wants them to explore different parts of the world, including her and her husband's home country, Ghana.

Although there have been challenges, such as trying to get her baby to swallow anti-malaria medication, bringing her kids on fieldwork trips has brought many rewards, says Dagadu, an adviser at Save the Children, a non-governmental organization in Washington DC. Her family has developed strong relationships with the caregivers in other countries who tended to Dagadu's children while she was in the field. And her children have come to feel a strong, personal link to their parents' native country. "They say they're from Africa, and they feel very much connected," says Dagadu. "That's turning out to be a big part of their identities."

Researcher-parents sometimes choose or might be obligated to take their children with them on fieldwork trips. These excursions often require extra planning, gear and creative thinking to incorporate childcare. But those who travel with their children say that their presence can help to break down cultural barriers and forge connections. Sometimes, those connections lead to insights that inspire fresh lines of research on topics such as local views on health care and cultural practices.

Children can also benefit from exposure to other cultures, as well as seeing their parents at work. Away from home, lifelong memories form that can shape families, says John Ward, an archaeologist at Lund University in Sweden. He and his wife, Maria Nilsson, who is also an archaeologist, spend months at a time doing fieldwork in Egypt, where they stay on a boat on the Nile River with their two children, aged four and one. "Life is short: children are part of that, and it's not a hindrance," Ward says.

CHILDCARE CHALLENGE

Taking kids into the field requires more planning than would a solo trip, and that preparation sometimes has to begin before your child is born, says Kelly Dombroski, a human geographer at the University of Canterbury in Christchurch, New Zealand. She ►

Archaeologist Colleen Morgan's daughter plays with a tool at a field site in Qatar.

PARENTING

The family way

Many researcher-parents are striving to conduct fieldwork while accompanied by their children.

► realized while she was pregnant that she would be taking her baby to China when he or she was about three months old. To prepare, she researched parenting styles that might make travelling with an infant easier.

As soon as her daughter was born in 2006, Dombroski used a sling to carry the baby, which eliminated the need for an unwieldy pram. “We really tried to think about how you travel with a baby,” she says.

Finding childcare solutions can be one of the toughest logistical challenges of bringing kids on research trips, especially because grants do not usually cover payments to caregivers. Some researchers can get help from family members, says Kathryn Grace, a geographer at the University of Minnesota, Twin Cities, who studies maternal and child health.

Grace did her PhD research in Guatemala when her first child was five years old. In the years that followed, she spent weeks at a time doing research in Burkina Faso in West Africa. In both countries, her schedule was unpredictable, and hiring a caregiver was logistically and financially impossible. Fortunately, her husband was able to join her.

Hiring caregivers is an option that can lead to the development of valuable relationships, and might be affordable, depending on the location. In Uganda, Dagadu paid US\$20 each day for two weeks’ of childcare. During a one-week-long trip to Senegal in 2017, she paid \$40 a day to a local colleague’s niece, who was on holiday from university, to mind both children. These amounts were fair wages for the caregivers, according to other local caregivers and Dagadu’s colleagues.

On her first visit to Uganda, Dagadu hired the cousin of a local colleague. When that caregiver graduated from university and got a job, the caregiver’s sister took over on a subsequent trip. After repeated visits to Uganda over several years, the sisters have become like family to Dagadu and her kids, who call them aunts.

Dagadu also researches the local norms to work out how much to pay caregivers, and includes potential caregivers in those conversations. These discussions enable her to meet the cultural expectations surrounding wages without paying exorbitant amounts that could make similar assistance difficult to arrange for other researchers in the field in the future.

Travelling with kids is often simple in the first few years of their lives, say researcher-parents. As kids get older, being creative with funds can help to make research trips work, Dombroski says. In spring 2018, she received a lump-sum fellowship that she used to support field research in Bhutan, Indonesia, China, Bangladesh, Australia and Italy. The money would have been enough to pay for a hotel and three meals a day in a restaurant. Instead, she used it during the Australia portion of the trip to rent a home and buy groceries so that she could accommodate her entire family, which now includes four children aged between 6 months and 13 years old. Her eldest daughter,

who was 11 at the time, also accompanied her to Bhutan.

Yet juggling finances for childcare doesn’t always work out. Dombroski was denied money from her university’s discretionary fund to bring her breastfeeding son to a writing symposium in Italy and therefore had to turn down the trip. And Grace took several months off from fieldwork in Burkina Faso in 2015 when her third baby was a few months old, because she didn’t want to take anti-malaria medication while she was nursing.

Grace recommends trying not to worry about missing opportunities for fieldwork or publishing papers when others are collecting the data — something with which she struggled at the time. Although her project — about climate, land use and health in West Africa — lost some momentum during her time away from the field, Grace was able to bring her son, then one year old, when she went back to Burkina Faso the following summer. As an assistant professor who was not yet up for tenure, she found that she could resume her research when she returned. “I thought at the time that I was really missing out,” she says. “But going back, I felt like I hit the ground running again.”

Shifting fieldwork closer to home can be another solution to accommodating childcare, says Christopher Lynn, an anthropologist at the University of Alabama in Tuscaloosa. He and his wife found out in 2002, while he was a master’s student, that they were having triplets. When the babies were born prematurely and needed extra care, he switched his research focus from topics that required far-away fieldwork to a study of Pentecostal churches near his home.

He now runs a laboratory that trains students to do local fieldwork. In a 2018 study, he interviewed more than 1,000 anthropologists and postgraduate students about how they balance fieldwork with looking after children

(C. D. Lynn *et al.* *PLoS ONE* **13**, e0203500; 2018). Most, he found, had family support to help subsidize their research. For scientists who don’t have independent resources or other ways to make it work, he advises looking locally. Sometimes, that might require a shift in research topic or a more flexible solution. “Some problems don’t have easy answers,” he says.

BABY ABROAD

Taking children on fieldwork trips is often not as idyllic as it might sound, and the experience can raise complicated emotions. When digital archaeologist Colleen Morgan’s daughter was eight months old in 2017, she and her husband, who is also an archaeologist, took the baby to their field site in Qatar. For six weeks, the family stayed in a large house with the rest of the field team. But the baby woke up throughout the night, sometimes every two hours. Some nights, amid much infant screaming, it took them hours to get her to sleep.

Morgan, a lecturer at the University of York, UK, brought along ear plugs, which she offered to everyone in the house. And although Morgan worried, the crying didn’t seem that loud to other team members, says Ben Sharp, a former field archaeologist now based in Billingshurst, UK, who lived in the house at the time. He was so tired by the end of the day, he says, that he slept through it all. And he appreciated the presence of a baby at the dig house. “I think that it was good for the team, having a baby around,” says Sharp, who is not a parent. “It was relatively amusing for everyone getting up in the morning and having breakfast with a baby. It gave us all a distraction from work when we needed it.”

However, Morgan still felt uncomfortable about imposing her life on her colleagues. She felt awkward about drying the parts of her breast pump in the house dish rack. And she worried about how the baby might be affecting

EASE THE WAY

Fieldwork checklist

Experienced researcher-parents who have taken children on fieldwork trips offer tips for smoothing potential bumps in the road.

- **Become acclimatized.** Colleen Morgan, a digital archaeologist at the University of York, UK, recommends finding out where the nearest hospital and medical-care providers are located before setting out.
- **Bring the right gear.** Depending on the child’s age, pack specific equipment such as a stout carrier and the longest-range baby monitor that is available. Morgan advises getting one that is battery powered — her own monitor enabled her to put her infant down to nap in other people’s houses, despite a lack of access to electricity.

- **Take sunscreen and a tent.** Many researchers bring powerful sunscreen; others take a small tent that offers some protection from ultraviolet light and mosquitoes so that a baby can nap inside.
- **Buy goods locally.** It’s usually possible — and can be a fun adventure — to acquire nappies, food and other necessities on site. Morgan points out that bringing fewer things reduces the risk of offending local people by implying that your gear is better than theirs.
- **Consider the potential costs of medical care.** Some researchers travel without health insurance and pay for visits to doctors. Others buy extra insurance to cover their children while travelling. **E.S.**



Archaeologist John Ward shows his daughter how to operate a camera in the field in Egypt.

team dynamics in an already intense situation. “You just feel like the mum that tried to push her own way in, or tried to do the impossible,” says Morgan. “That was really a struggle.”

The experience can also raise parenting challenges, says Grace, whose five-year old daughter attended a school in Guatemala without any Spanish language skills when they arrived in 2006. She eventually became fluent in Spanish, made friends and got invited to classmates’ birthday parties. But it was challenging for the first couple of months, and, Grace didn’t feel comfortable asking for help. She didn’t even know who to ask. “I felt like it was my fault for having a kid, and so I needed to deal with that, like, I didn’t want to burden anyone,” she says. “I was working a lot and she was struggling and we didn’t always have the support that we needed.”

Researchers who have juggled family life with fieldwork recommend considering health and travel insurance, and seeking support from supervisors, colleagues and other researchers who have had similar experiences. Until April, Dombroski belonged to a private social-media group of ethnographers that hosts discussions about doing fieldwork while accompanied by kids. Members pose thorny questions and offer their experience or potential solutions, Dombroski says.

Health insurance is another concern, and researchers can take various approaches to planning for illness and travel emergencies. University insurance often covers researchers but not their families (see ‘Fieldwork checklist’).

To get work done with children around, experienced researcher-parents suggest altering work schedules. In Qatar, Morgan got up at 4.30 a.m. to work before her baby awoke. On days that she went into the field, she would take the baby with her until midday, when temperatures rose. After her husband returned from the dig later in the afternoon,

she would hand over the baby and resume work at her computer.

As children get older, their care needs change, Morgan points out, and field researchers will need to adapt. At eight months old, she says, her daughter’s sleeping habits were unpredictable, but she was portable. On Morgan’s next trip to the same location, when her daughter was 20 months old, the baby’s naps were longer and more regular, but she demanded more attention and couldn’t just be carried around to accommodate work tasks.

By the time kids reach school age, Ward adds, parents might have to make some tough decisions. He and Nilsson will need to work out whether they will take turns to go to Egypt for fieldwork, or both go for shorter periods, so that their daughter can attend kindergarten in Sweden.

Being flexible during a fieldwork trip is key, says Dombroski. She returned to Asia for a one-year stay in 2007 to work on her PhD project, when her daughter was eight months old. Her original goal had been to interview Muslim women about the local economy in a village that required a drive over a high-altitude pass to reach. But the elevation might have been too high for the baby, who wouldn’t have been able to say if she was feeling sick. And Dombroski couldn’t leave the infant, who was still breastfeeding, at home. Instead, she worked with her supervisor to develop another project, which focused on women from all ethnic groups in a city lower down on the Tibetan Plateau.

BENEFICIAL FOR ALL

Despite the logistical and financial struggles, seasoned researcher-parents say that the rewards of bringing children into the field are many, and can even include the development of new research questions. In China, Dombroski found that the presence of her baby prompted conversations and unsolicited advice, which

caused her to question her assumptions about parenting. She received disapproving remarks, for example, about her baby’s outerwear, an issue that she had never considered when living in New Zealand, where children wear short trousers all year round, even in winter.

The comments led Dombroski to investigate the idea of ‘awkward engagements’, in which the cultural behaviours of people from differing backgrounds don’t match. That work led to several publications that resonated with other researchers. Some of her most cited work, she says, has been the result of research trips that included her children, which suggests that their presence did not inhibit her work. And she regularly gets e-mails from other researchers about those papers and her blog, *Thowntogetherness*, which includes articles on parenting. Family, she says, “doesn’t have to be something that holds me back”.

Similarly, Grace published a study on the links between climate change and reproductive health (K. Grace *Nature Clim. Change* 7, 479–485; 2017), which was influenced by her own experiences of breastfeeding, as well as talking to women about breastfeeding in Burkina Faso.

Grace thinks that sharing stories of including children in fieldwork with other scientists is an important step towards changing funding policies to help parent-researchers to bring their kids into the field, and can give researchers the confidence to do so, she adds.

“Science benefits from having women willing to go into the field while nursing their babies,” she adds.

“Science benefits from having women willing to go into the field while nursing their babies, and I think it’s a shame that the costs are on us,” Grace says.

She has started to add a note to her syllabi to say that babies are welcome in class. The challenges she faced when first taking her kids into the field “made me feel like I was doing it wrong”, she says. “But I think science has it wrong.”

Ward marvels at his four-year-old daughter, who can speak English, Swedish, Arabic and some French, thanks to her exposure to a multinational research team. His son, who turned one while they were in Egypt, thinks of the research boat as a big fort. Both kids have become accustomed to finding 3,500-year-old human remains spread out on the dining room table.

“It’s a totally different life and I wouldn’t change a thing,” Ward says. “There is a euphoria in this family unit of ours. It’s wonderful and I don’t think you can easily put that into words, no matter all the ups and downs. If you have the opportunity to do it, do it. You’ll never regret it.” ■

Emily Sohn is a freelance journalist in Minneapolis, Minnesota.

THE MEMORY LANTERNS OF LOI KRATHONG

Something to share.

BY PRESTON GRASSMANN

Today is a holiday for purging bad memories. Until now, the sky lanterns and floating candles have largely been symbolic, but new techniques in selective memory-erasure have changed all that.

Chakkri carries his new lanterns from Sukhumvit's Atlanta Hotel, all the way to this smoky blur of neon at the edge of Chao Phraya River.

A crowd gathers around him, eager to buy the festive paraphernalia of Loi Krathong.

With his lanterns displayed, Chakkri holds one of them out for the gathered buyers to see. He channels through a sequence of tailor-made images and the crowd steps back — countless faces flashing in reflection. Stock images of news clips and footage of war scenes explode across the pixelated surface of the paper. He lights a lantern and lets it go — a gust of wind pushing it out over the stalls of Canvas Town and over Chao Phraya, where it joins other lanterns ablaze with memory. Young and old stand along the shores of the river, watching their own selected moments of time drift away.

As his hands pass swiftly over the counter, exchanging lanterns for old baht, Chakkri notices a boy in the crowd, pushing his way forward.

He has seldom seen such eagerness in a child's eyes, and for a moment it reminds him of his own childhood memories of Loi Krathong, the amazement he had felt when his mother had taken him to the festival so many years ago.

As the boy reaches the stall, he looks up at the images and videos that flash across the lantern screens. "Are those your own memories?"

Chakkri shakes his head and leans over with a conspiratorial smile. "I'm afraid not. They're modified from the public domain."

"Don't you have any memories of your own you want to ...?"

"Erase?" Chakkri says, shaking his head. "I forget enough as it is."

The boy looks up. Even through the flickering wash of lantern light, Chakkri can see the



disappointment. For a moment, it catches him off-guard.

A man steps forward to stand behind the boy, placing a hand on his shoulder. The boy doesn't turn back.

"Even the painful memories?" the boy asks.

"Especially the painful ones," Chakkri says. "I'm not sure who I would be without them."

"I'm sorry," the man says, shaking his head sadly. "My son ... lost his mother a few days ago ..."

Chakkri holds his hands together and closes his eyes, memories of his own mother returning as he offers his condolences.

The boy traces a finger over the surface of a lantern, as if the words are too much for him to hear, as if he's trying to find his way out of a maze of pain.

Chakkri lifts a hand to the crowd, signalling that he's closing the booth. The boy's lantern remains lit, while the others fade to blank screens and he leans towards the boy again.

"We can share the past, instead of forgetting it," he says. Appearing on the lantern's screen is a sky full of lights — a bright mosaic of his own memories floating in the dark. It might've been that very night along the Chao Phraya, but this was 30 years ago in another city. He lets it play for a moment before revealing a woman and a child.

"Is that ...?" the boy asks.

"My mother," Chakkri says. "And me."

He stands on the shores of the Chiang Mai River, holding his mother's hand as they watch the lanterns of Loi Krathong — a memory he has never shared.

"Now it's your turn."

He places the lantern in the boy's outstretched hand. He is too young to have a neural implant of his own, so his father takes a moment and selects a memory.

A video sequence begins to play on the lantern's screen — a child at a beach, crying as the waves surge and fall against white sand. The mother lifts the child in her arms, pointing out across the sun-burnished sea, where tiny islands and ships dot the horizon. The child's fear begins to fade as he takes notice of faraway places he has never seen before. He leaps from his mother's

arms and runs along the shore on his own, unafraid. "I remember that," the boy says, looking back at his father. "I remember that."

When the sequence ends, Chakkri leans down to face the boy at eye level.

"That's a beautiful memory," he says, pointing at the lantern. "Why don't you keep that?"

An echo of that childhood memory rises back into the boy's face now, and other emotions take the place of pain. The boy's eyes shine through his tears, and the father places his hands together, thanking Chakkri for his kindness.

For a moment, Chakkri is a child again, holding his mother's hand, the blazing lanterns above like distant islands, like ships setting sail to some place faraway. And then he remembers the words she had said that day, all those years ago: "The hardest part of holding on to memories is the loneliness of never sharing them."

As he watches father and son walk away, holding a memory in their hands, he realizes how right she had been. ■

Preston Grassmann is a contributing editor of *Locus Magazine*, where he writes a regular column called *The Cosmic Village*. His recent work has been published in *Shoreline of Infinity*, *Daily Science Fiction* and *Mythic Delirium*. He currently lives in Tokyo.

ILLUSTRATION BY JACEY

nature INDEX 2019 ANNUAL TABLES

NATURE, VOL. 570, ISSUE NO. 7761 (20 JUNE 2019)

BEHIND THE NUMBERS

We are delighted to present the Nature Index Annual Tables. The rankings for 2018 revealed in this print and online supplement highlight the institutions and countries with the highest outputs of top-quality research in the natural sciences. Our measure, fractional count (FC), is based on the share of articles published in 82 prestigious scientific journals, selected by an independent panel of scientists and tracked by the Nature Index database. The fractional count is more fully explained on S3.

Although we rank institutions by comparing their FC, we recognize there are many qualities that distinguish them. There are also many possible causes of a decrease in an institution's publications in Nature Index journals from year to year. Yet, the specificity of our metric is also its strength: top-notch research in the natural sciences, pure and simple.

Renowned institutions — often with advantages conferred by long history, good reputation, generous funding, and size — reign over the research ranks as might be expected. It is interesting, then, to see how the rankings change when we consider an institution's Nature Index output as a proportion of its overall output in the natural sciences, as a way of seeing which institutions might be punching above, or below, their weight.

This normalized ranking, showing what share of an institution's research output in the natural sciences has been judged high quality, reveals a very different set of leaders among academic institutions. Some of them have tiny



The National Institutes of Health, ranked 2nd in the life sciences, has the largest zebrafish genome library in the United States.

article counts compared to the likes of Harvard: the Cold Spring Harbor Laboratory in New York is number one, yet comes 345th in our Annual Tables; the 30-year-old Jawaharlal Nehru Centre for Advanced Scientific Research of Bangalore, India, is 8th in the normalized ranking and otherwise 413th. Princeton University is an impressive 9th normalized and 24th otherwise. Some academic institutions fare much worse than their reputations might predict, and others, much better. As the normalization analysis demonstrates, size isn't everything, and neither is rank.

For more detail, more tables and more top 10 rankings, please visit www.natureindex.com.

Catherine Armitage
Chief editor, Nature Index



ON THE COVER

The Nature Index 2019 Annual Tables reveal the dominant institutions in natural sciences research, according to their output in leading journals.

EDITORIAL: Catherine Armitage, Gemma Conroy, Bec Crew, David Payne, Stephen Pincock, Rebecca Dargie **ANALYSIS:** Bo Wu, Willem Sijp **ART & DESIGN:** Tanner Maxwell, Chika Takeda, Annthea Lewis, Wojtek Urbanek **PRODUCTION:** Kay Lewis, Ian Pope, Nick Bruni, Bob Edenbach, Joern Ishikawa **MARKETING and PR:** Stacy Best Ruel, Angelica Sarne, Elizabeth Hawkins **SALES & PARTNER CONTENT:** Marcos Valiente, Sabrina Ma, Yingying Zhou, Ruffi Lu **PUBLISHING:** Rebecca Jones, Richard Hughes, David Swinbanks.

NATURE INDEX 2019 ANNUAL TABLES
Nature Index 2019 Annual Tables, a supplement to *Nature*, is produced by Nature Research, the flagship science portfolio of Springer Nature. This publication is based on data from the Nature Index, a Nature Research website maintained and made freely available at natureindex.com.

NATURE EDITORIAL OFFICES
The Campus, 4 Crinan Street,
London N1 9XW, UK
Tel: +44 (0)20 7833 4000
Fax: +44 (0)20 7843 4596/7

CUSTOMER SERVICES

To advertise with the Nature Index, please visit natureindex.com/ clientservicesfeedback@nature.com
Copyright © 2019 Macmillan Publishers Limited, part of Springer Nature.
All rights reserved.

Global leaders

S2: The US reigns, but China is taking up ever more space, squeezing out European stalwarts.

World of research

S4: While the Nature Index top 20 also dominate the subject rankings, outliers offer surprises.

Top 100 table

S5: Institutions ranked by their Fractional Count in the Nature Index.

GLOBAL LEADERS

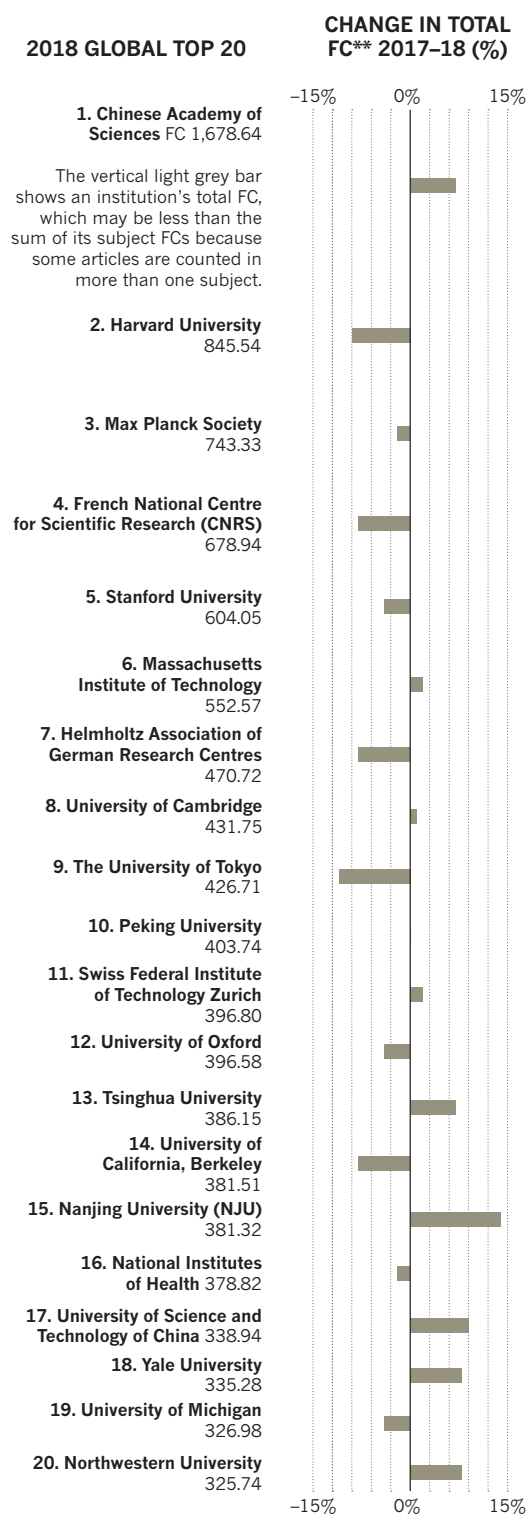
While the Nature Index top 20 institutions also dominate our 2018 subject rankings, outliers offer some surprises.

SOURCE: NATURE INDEX

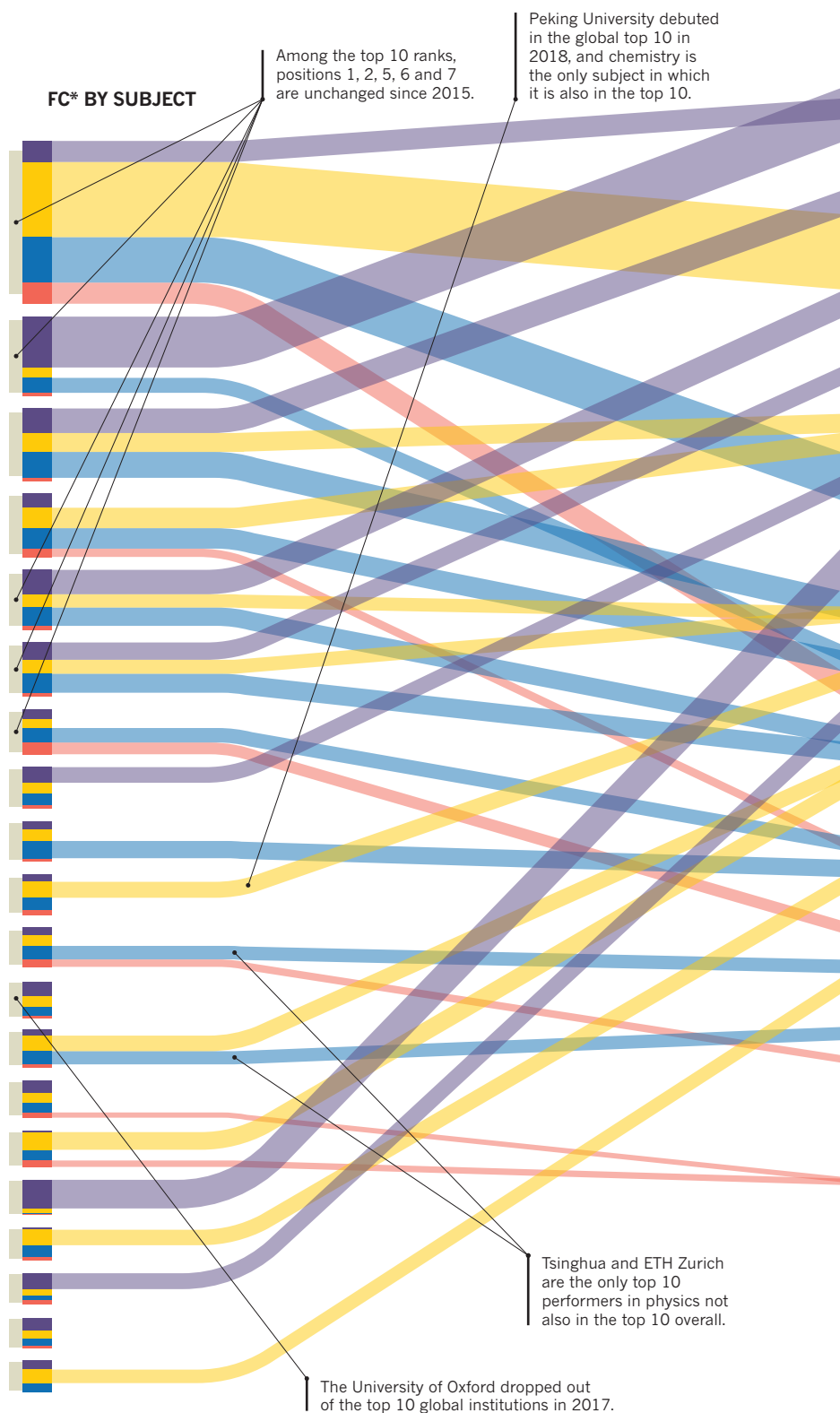
THIS GRAPH EXPLAINED

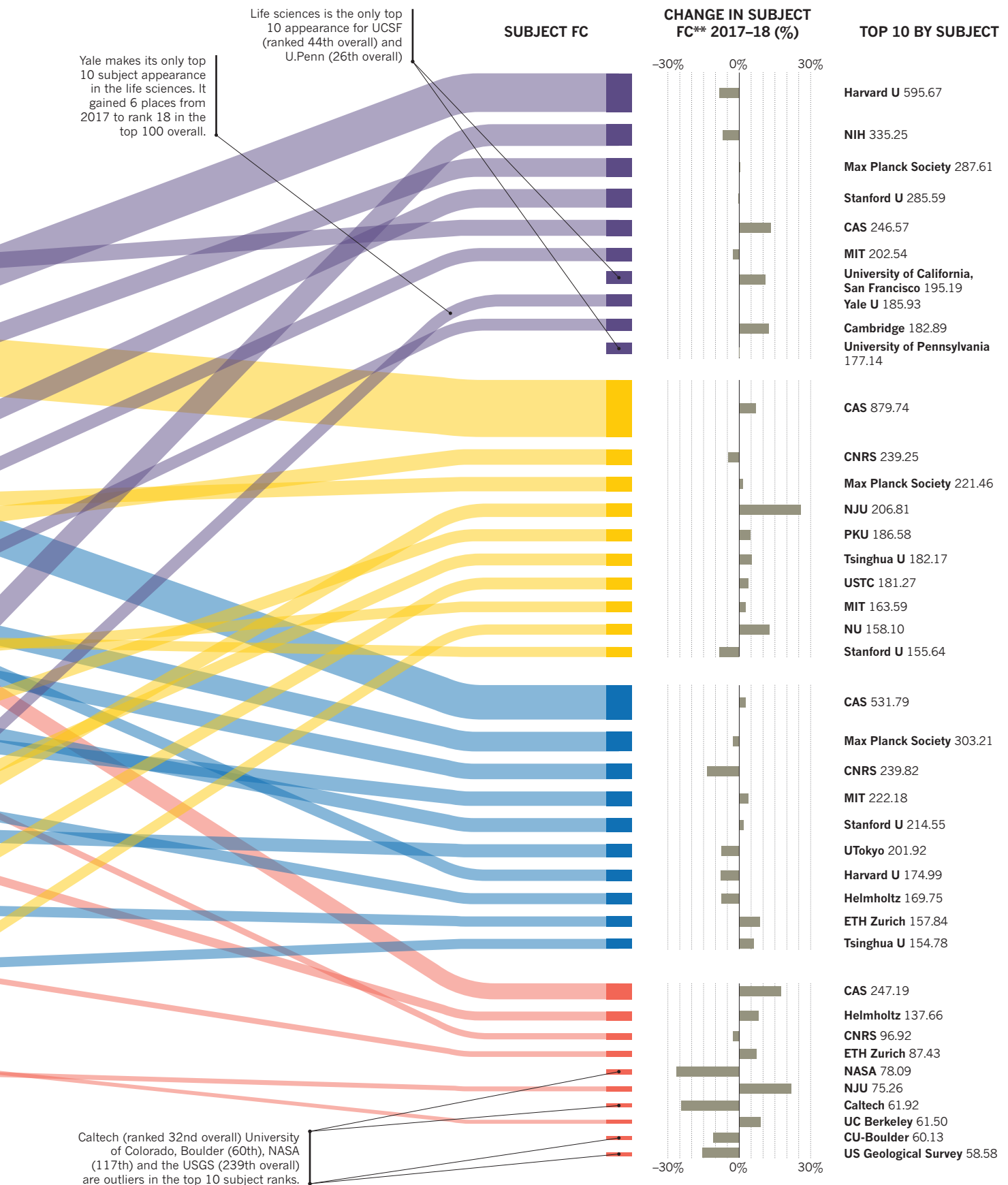
The left axis of the central graph shows the top 20 global institutions ranked by Fractional Count (FC)*. The right side shows the top 10 institutions in each of the four natural sciences subjects covered by the Nature Index ranked by their FC in that subject. Line widths are proportional to each institution's FC, and the vertical bars on either side show the percentage change in FC.

Overall Life sciences Chemistry Physical sciences
Earth & environmental sciences



FC* BY SUBJECT





*Fractional Count (FC) is assigned to institutions based on the contributions of their affiliated authors to articles published in the 82 journals tracked by the Nature Index database, with all authors on each article considered to have contributed equally, and a maximum combined FC for any article of 1.0.

**Here FC is adjusted to 2018 levels to account for a small annual variation in the total number of articles published in the journals.

WORLD OF RESEARCH

The US reigns as the colossus but China is taking up ever more space, squeezing out European stalwarts.

